



# Data 2012

- Sign up for Thursday dinner excursions by end of break today at 11:15 (menus at desk outside this room, Room 137).
- Return paperwork in your folders to Robert Ping by 5:30 p.m. today (if there was paperwork in your folder at registration).
- Feel free to hang your Posters from last night on the walls in this room or BoF rooms.
- Room numbers listed for Lightning Talks and Featured Talks on pages 5-6 should be this room, Room 137 (the agenda on pages 2-3 is correct)

**Have a great meeting!**

# DataNet/INTEROP: Advances, Collaborative Interactions, and Germinating for Long Term

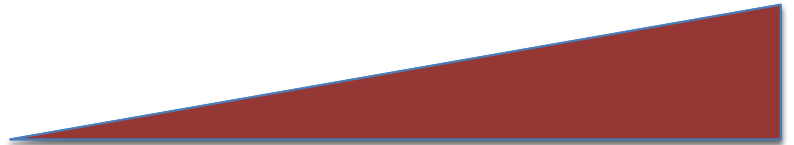
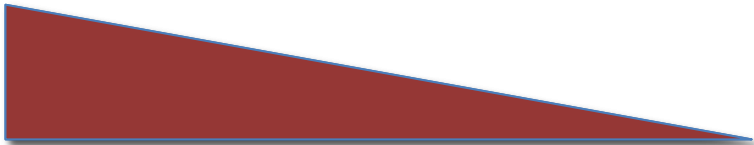
Beth Plale

Professor, School of Informatics and Computing

Director, Data To Insight Center

Managing Director, Pervasive Technology Institute

DataNet/INTEROP Workshop organizer



# DataNet / INTEROP

- NSF DataNet - Sustainable Digital Data Preservation and Access Network Partners - exemplar national and global data research infrastructure organizations (dubbed DataNet Partners) that provide unique opportunities to communities of researchers to advance science and/or engineering research and learning
- NSF Interop - crosscutting program supports community efforts to provide for broad interoperability through the development of mechanisms such as robust data and metadata conventions, ontologies, and taxonomies. Responsible for consensus-building activities and for providing the expertise necessary to turn the consensus into technical standards.
- NSF OCI Task Force on Data and Visualization report, final draft, March 7, 2011



# Purpose

- Hear from the DataNet and INTEROP projects
- Collective expertise in researchers engaged in DataNet and INTEROP is tremendous.
- Information exchange between DataNets and INTEROPs

*→ Stretch goal of workshop is to plant first seeds for long lived, collective effort that advances scientific sharing and preservation worldwide*

# Roadmap

- Information exchange
- Interaction between projects: war stories, how to contribute, identifying targets for incorporating results (from INTEROP -> DataNet)
- Creating something greater than sum of parts that aligns with EU partners and beyond

Posters, lightning talks.  
This morning.

**Birds Of Feather:**  
informal conversations  
on topic of common  
interest.

**Educational:** hearing  
from IETF key  
founders; EU EUDAT  
and DAITF efforts

# EarthCube Community Charrette Nov '11:

## 14 Key Capability Categories emerged

Dataset and workflow discovery	Data access services	Tools to Probe, Validate, Verify, and Visualize Data	Data security and trust
Workflow execution management	Data management within workflows	Metadata for workflow and data sets	
Numerical methods and software engineering	Modeling standards and frameworks	Modeling capabilities within Cloud, Grid, HPC, and Science Portals	
Policy enforcement processes	Best practices & Governance Models for definitions & standards	Broad Participation: NGO, industry, domain partners, international	Continuity, Sustainability, & Evolution

# Rules of Engagement

- Topics identified as we go; emerge by participant preference.  
Suggestions for today's BOF
  - Storage solutions and funding models for long term storage
  - DataCite, its use
  - IETF as model to continue momentum after workshop
  - Issues in using cloud storage for long term availability
  - Dataset discovery and data access services
  - Broadening participation
- Inclusion principle: on stretch goal there are no predestined players. We're trying to build something part grass roots, part top down. It'll take a lot of work by volunteers committed to vision of alliance for advancing scientific preservation. We're looking for committed people to join us.

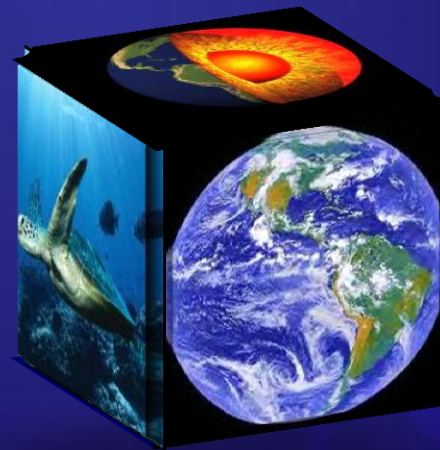
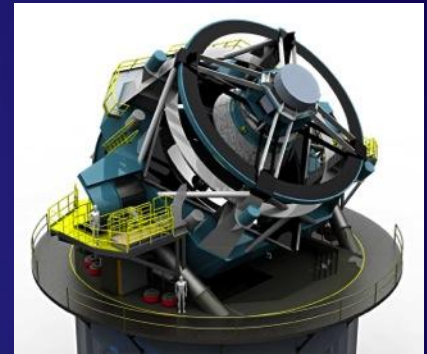
# Scientific Data: Changing and Changing The World

Rob Pennington

Program Director

Office of Cyberinfrastructure

NSF



# The Shift Towards a "Sea of Data"

## *Implications*



### ❖ All science is moving towards a data-dominated world

- Fundamental questions become focused around data: How to remove boundaries? How to incentivize sharing?

### ❖ Software

### ➤ Publications

How do we attribute credit for this new publication form? How are data peer reviewed? What is a publication in the modern data-rich world?

### ❖ Totally new methodologies

- Algorithms, mathematics, culture

### ❖ Data become the medium for

- Multidisciplinarity, communication, publication...science



# Recent Data Related Activities

- ❖ National Science Board Task Force on Data Policies Recommendations
  - [http://www.nsf.gov/nsb/committees/tskforce\\_dp.jsp](http://www.nsf.gov/nsb/committees/tskforce_dp.jsp)
- ❖ US Office of Science & Technology Policy
  - <http://www.whitehouse.gov/blog/2011/11/07/request-information-public-access-digital-data-and-scientific-publications>
    - “Request for Information on Public Access to Digital Data and Scientific Publications”
    - “Public Access to Digital Data Resulting From Federally Funded Scientific Research”
- ❖ Coordinated US-EU programs
  - INFRA 2012-3.1, 3.2 and OCI/GEO/MPS DCL (STCI)<sup>6</sup>



# Changes Coming at NSF for Data!

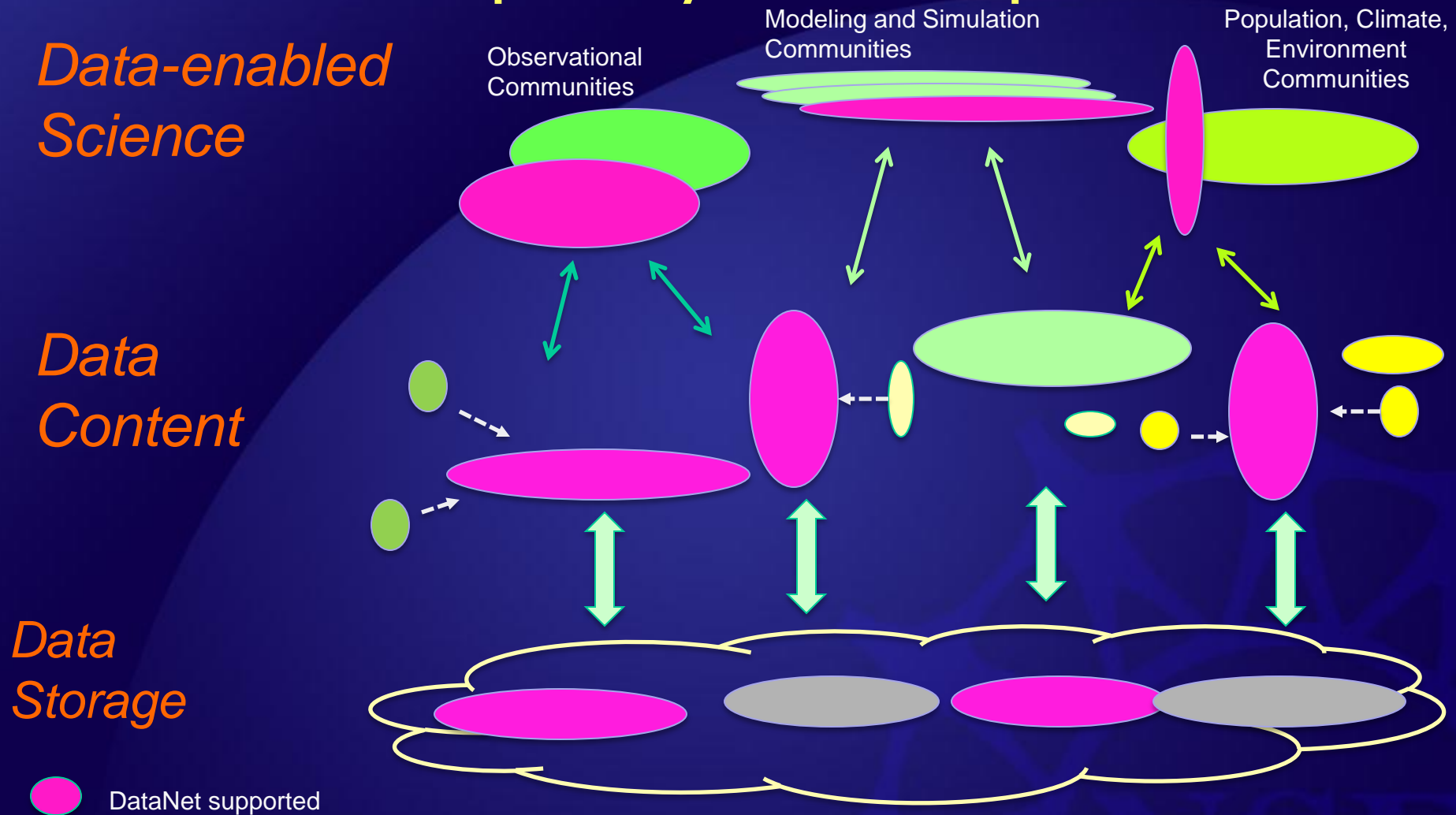
- ❖ Data are becoming:
  - Primary means of communication through sharing
  - Major product of research (including publication)
- ❖ Long-standing NSF Policy on Data:
  - *"Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data... created or gathered in the course of work under NSF grants"*
- ❖ NSF now requires a Data Management Plan (DMP):
  - 2-page supplement to the proposal
    - Subject to peer review; criterion for award
  - Not possible to submit proposals without DMP



# Data-Enabled Science

- ❖ Data Services Program (*data*)
  - Provide reliable digital preservation, access, integration, and analysis capabilities for science and/or engineering data over a decades-long timeline
- ❖ Data Analysis and Tools Program (*information*)
  - Data mining, manipulation, modeling, visualization, decision-making systems
- ❖ Data-intensive Science Program (*knowledge*)
  - Intensive disciplinary efforts, multi-disciplinary discovery and innovation

# DataNet: A Multi-tiered and Multi-Disciplinary Landscape



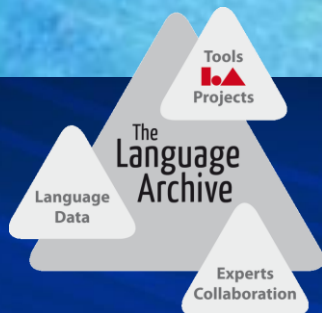
# Summary

- ❖ Think about effective ways to approach the challenges associated with data
  - Critical concepts and goals
  - Realistic and innovative
- ❖ Structure for longevity
  - Scalable open inclusive governance
  - Long term business models

# **RIDING THE WAVE**

**FROM VISIONS TO ACTIONS FOSTERING  
EUROPEAN RESEARCH  
HLEG ON SCIENTIFIC DATA**

**- AN ACTION FROM THE EUROPEAN  
COMMISSION -**



Peter Wittenburg  
The Language Archive - Max Planck Institute for Psycholinguistics  
Nijmegen, The Netherlands



- ❑ **Group and Motivation**
- ❑ The Research Data World
- ❑ Opportunities and Challenges
- ❑ Collaborative Data Infrastructure
- ❑ Relevant Aspects
- ❑ Vision 2030
- ❑ Action Points

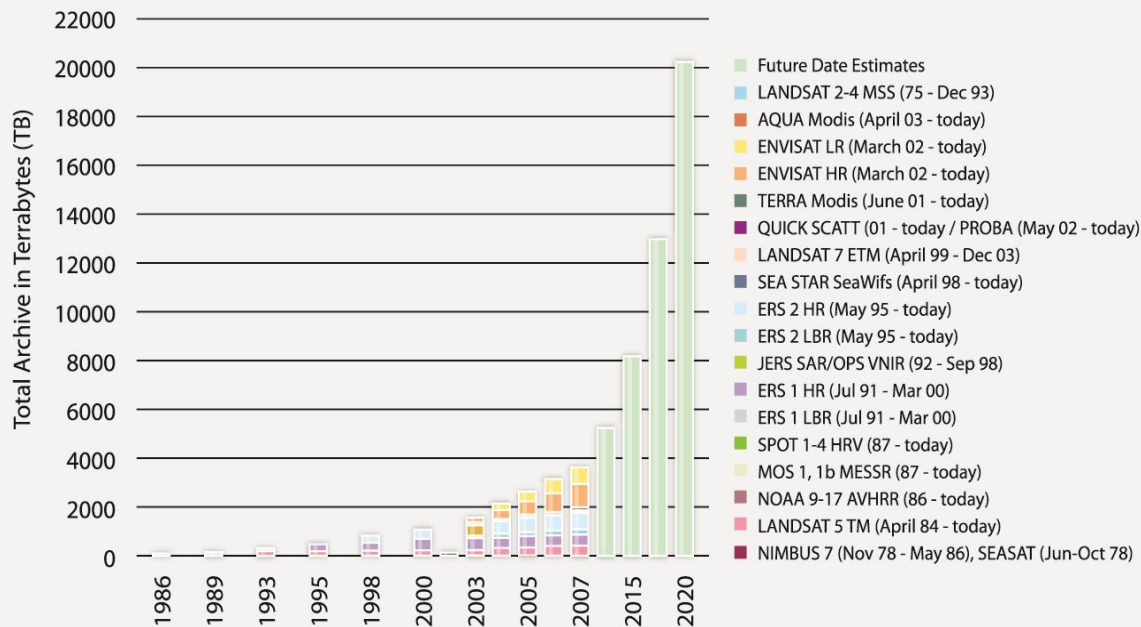
# HLEG and Motivation

- EC DG InfSo invited 11 experts and 12 contributors to 6 meetings  
**Chair:** John Wood; **Rapporteur:** David Giaretta; **Members:** Thomas Andersson ; Achim Bachem; Christoph Best ; Françoise Genova ; Diego R. Lopez; Wouter Los; Monica Marinucci; Laurent Romary; Herbert Van de Sompel ; Jens Vigen; Peter Wittenburg; **EC Representatives:** Konstantinos Glinos; Carlos Morais-Pires;
- Goals
  - come up with a vision 2030 for the management of research data as a guideline for future actions of the EC
  - discuss all relevant aspects around “data” in an unbiased manner
  - accelerate measures to take care of our data and to remain competitive
- Motivation
  - enormous increase in scale and complexity
  - not only summarize what some of us already know or are doing, but facilitate a systematic and global approach and push ahead actions
  - knowledge is power - data has a value although difficult to quantify

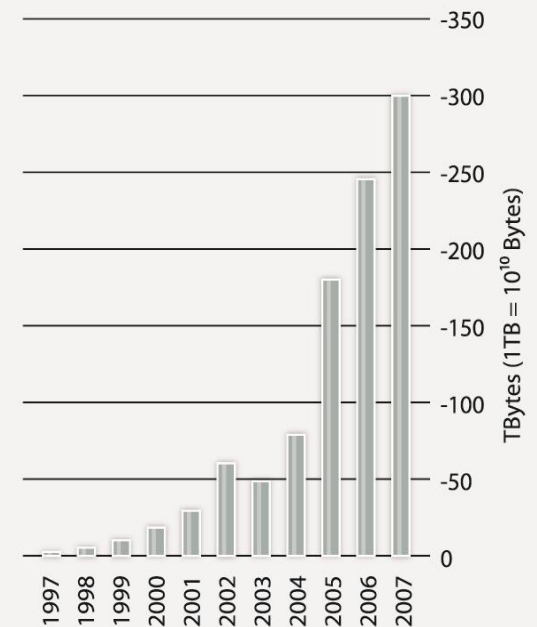
# Trends are Known

“A fundamental characteristic of our age is the raising tide of data – **global, diverse, valuable and complex.** In the realm of science, this is both an **opportunity** and a **challenge.**”

Evolution of ESA's EO Data Archives between 1986-2007 and future estimates (up to 2020)



Yearly Data Creation on NICE

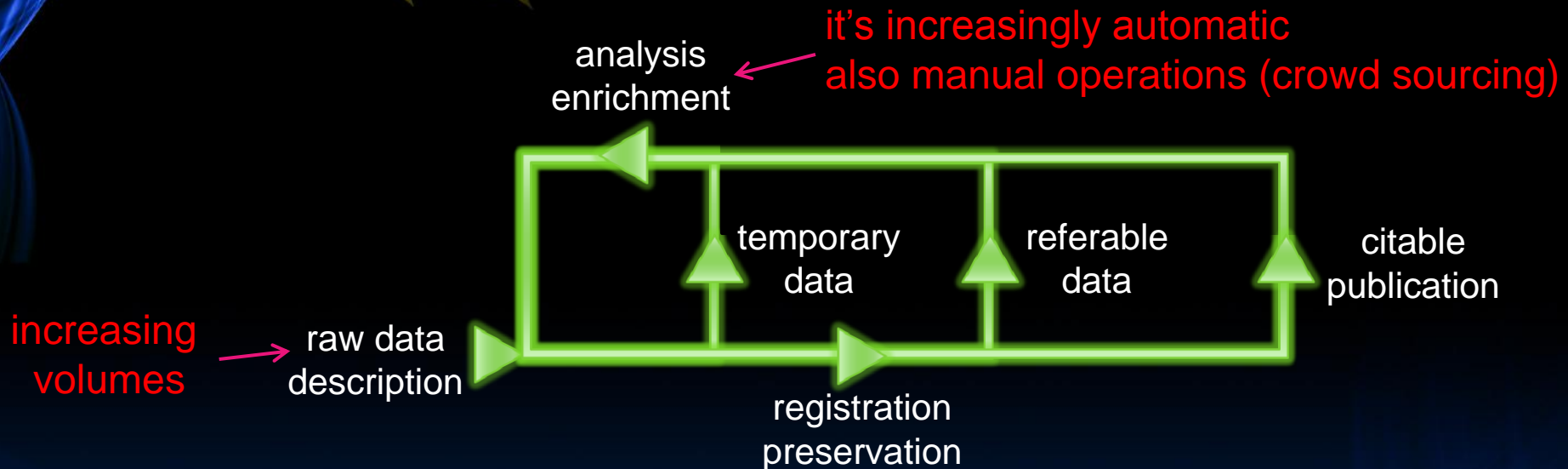


- 
- ❑ Group and Motivation
  - ❑ **The Research Data World**
  - ❑ Opportunities and Challenges
  - ❑ Collaborative Data Infrastructure
  - ❑ Relevant Aspects
  - ❑ Vision 2030
  - ❑ Action Points



# Research Data World

- Knowledge Creation Cycle is changing - almost all disciplines



- Exabyte scale and millions of related files of different types create unseen complexity - **deal with a new quality**
- much relevant data is and will not be registered  
(80 % of recordings about languages and cultures are endangered)

# Berman's classification

## The data pyramid - a hierarchy of rising value and permanence

### Digital Data Collections

Reference, nationally and internationally important, irreplaceable data collections

Key research and community data collections

Personal data collections

Increasing constituency

Increasing value

Increasing trust

Societal Value  
**Patrimonial Data**

Community Value  
**Cyclic Data**

Individual Value  
**Transient Data**

decreasing risk of loss or damage

Increasing responsibility

Increasing stability

Increasing infrastructure

### Respositories/ Facilities

National- and international-scale respositories, libraries, archives

"Regional" - scale libraries and targeted data archives and centers

Private respositories

Source: Adapted from Francine Berman, UC San Diego, in *Communications of the ACM*.

this is the data we need to take care of but do we know which data will be of relevance for future generations?

- some interesting aspects
  - lossless **separation of content and carrier** in the digital domain changes the world - some speak about a revolution comparable with the invention of book printing
  - **data creators are not known personally** to data users anymore - we need to solve the trust problem
  - research world is one of the **primary driver** for the data tides
  - there is no doubt: **data accessibility** changes nature, pace and direction of research
  - **diversity** in many dimensions is the dominant feature of scientific information and this will probably increase due to the inherent innovation forces
  - technology allows to **include the citizens** in different roles - also as contributors, increasing volume and complexity
  - increasing pressure towards **open access**

- 
- ❑ Group and Motivation
  - ❑ The Research Data World
  - ❑ **Opportunities and Challenges**
  - ❑ Collaborative Data Infrastructure
  - ❑ Relevant Aspects
  - ❑ Vision 2030
  - ❑ Action Points

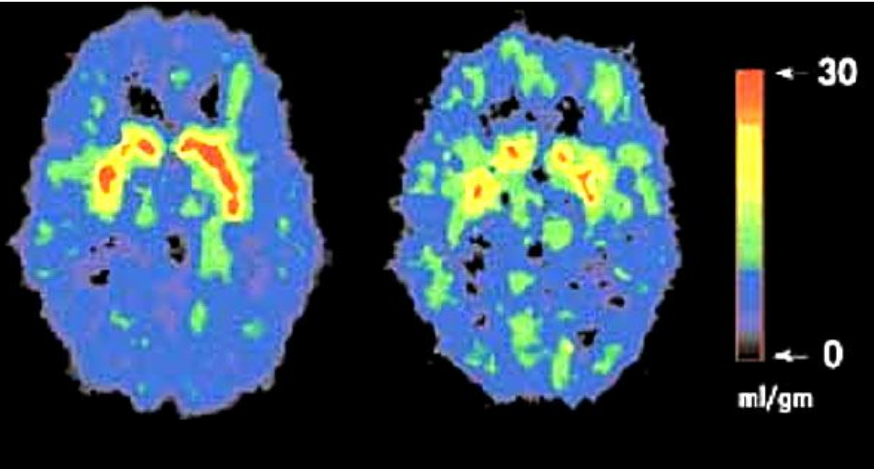
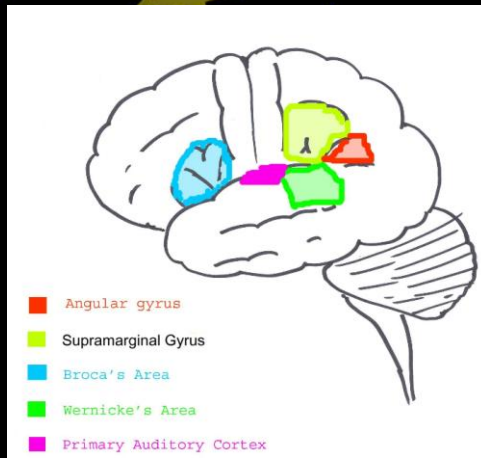


# Opportunities and Challenges

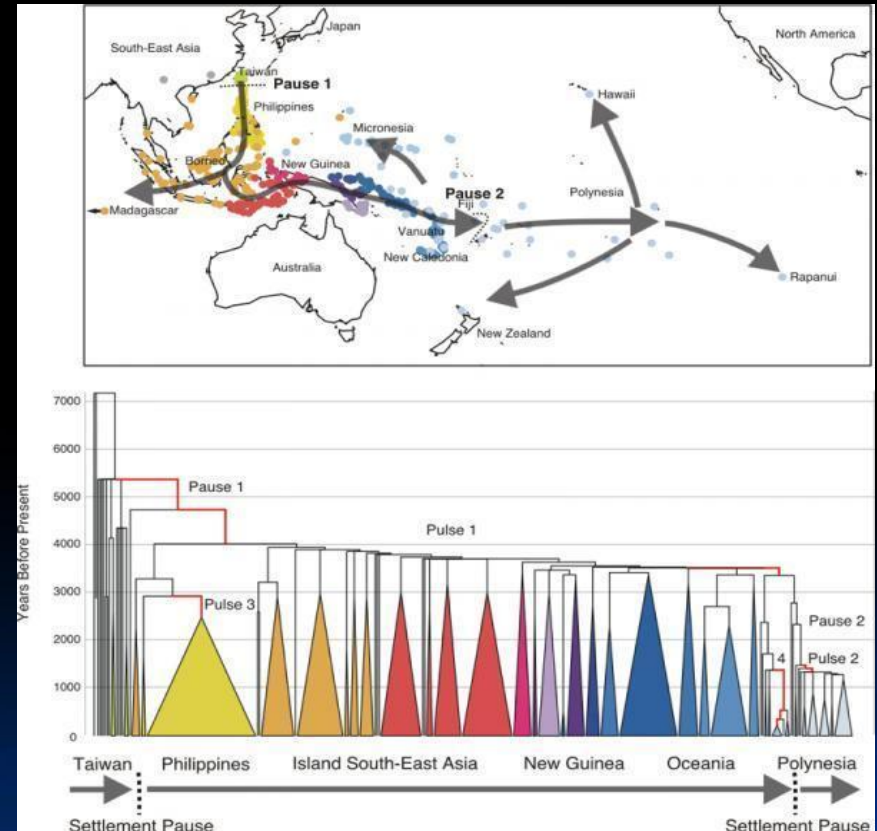
- virtual integration
  - **integrating large data** sets across disciplines and countries to create new insights
  - recombining data to **virtual collections** from different perspectives
  - sufficient data as basis for **comprehensive modeling** and understanding
  - **data intensive science**: find correlations and draw inferences not constraint by pre-assumptions - huge amounts of data not used
- tackling the **grand challenges** resulting from human activities
  - climate change, sustainable energy, stability health, etc.
  - **stability of our societies and minds** given the innovation, changes, globalization and migration
- facilitating the many “**small research questions**” driven by scientific curiosity
- relieve researchers from **data management and curation** effort (40% of knowledge workers time spent on finding and transforming data)

# Opportunities and Challenges

two examples for “small research questions” from my domain:  
languages and language processing



finding more about the functional  
architecture of our brain



finding more about the roots of  
our languages and cultures

# Opportunities and Challenges

- however there are quite some hurdles to overcome
  - need to **change culture** and researchers minds to deposit data
  - need to establish **trust** at depositor's and user's side
  - trust has to do with **data quality, integrity and authenticity**
  - need to convey **context and provenance** to allow users to understand
  - need **new responsibilities** and **new mechanisms** to solve data curation, preservation, organization and granting access without ignoring security and ownership principles
  - need **incentives** for researchers to deposit in proper quality so that data publication helps in career and reputation building

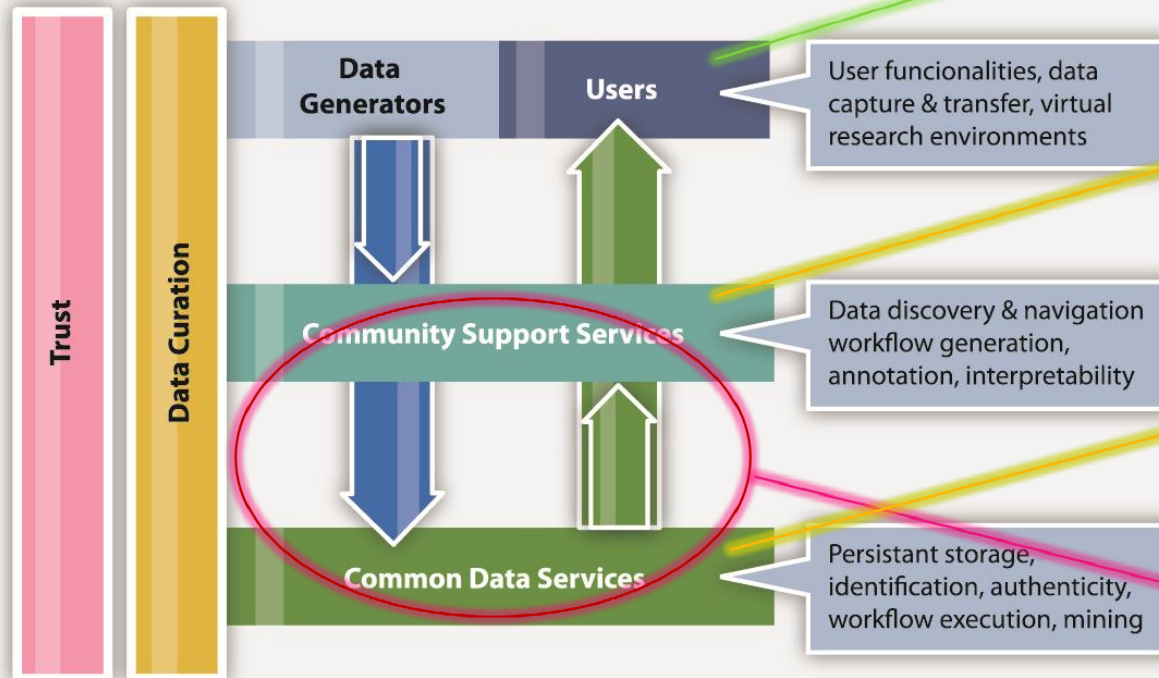
- 
- ❑ Group and Motivation
  - ❑ The Research Data World
  - ❑ Opportunities and Challenges
  - ❑ **Collaborative Data Infrastructure**
  - ❑ Relevant Aspects
  - ❑ Vision 2030
  - ❑ Action Points



# Collaborative Data Infrastructure

- obviously we need a new layer of responsibility:  
a **systematically constructed and global data infrastructure**
- some already working on data organizations - piecemeal, fragmented
- we call it a **Collaborative Data Infrastructure** open for many players and heterogeneity based on an abstract architecture and proper APIs

## The Collaborative Data Infrastructure - a framework for the future



many researchers from different disciplines and with different interests

CLARIN, CESSDA, DARIAH, LifeWatch, ENES, EPOS, etc.  
>40 R Infrastructures

EUDAT, OpenAIRE, D4Science, etc.

are faced with large heterogeneity  
need an architecture

- speaking about abstractions
  - data **object level** vs. data **content level**
    - can put cherries, apples, potatoes, etc. all in the same container
    - the way to treat them for making food is different
  - data object architecture is common (?)
    - early Internet discussion: is **email** specific for a discipline?
    - **PID** to identify, to ensure integrity and authenticity, etc.
    - **metadata** to describe context and provenance
    - how to treat **collections** (ORE, etc.)
    - have **many instances** (copies) at various locations (preservation, etc.)
    - **interoperability** at object level is about MD, PIDs etc.
  - data content level is discipline specific
    - it's about **structure** (schemas) and encoding schemes (MPEGx, etc.)
    - it's about **semantics** (vocabularies)
    - **interoperability** is discipline and **USAGE** specific

# Collaborative Data Infrastructure

coming there is a step-wise and layered process requiring some time and proper abstractions - but needs to be driven bottom-up

don't have to start from scratch

but success depends on careful, coordinated and agile planning

don't have to start from scratch

Separate disciplines



- 
- ❑ Group and Motivation
  - ❑ The Research Data World
  - ❑ Opportunities and Challenges
  - ❑ Collaborative Data Infrastructure
  - ❑ **Relevant Aspects**
  - ❑ Vision 2030
  - ❑ Action Points

# Relevant Aspects

- funding
  - need to understand data as a socio-economic treasure in a competitive domain - at the end research is about **global competition**
  - need proper **business models** - who is paying, which data is free, etc.
  - **governments** will have to reserve funds for data management
- quality and impact
  - need to measure quality and impact, which **metrics** are meaningful
  - need to **reward** contributors but how?
- management/curation skills
  - need a new type of experts: **data scientists**
- power researchers
  - resulting CDI will be complex as the data world will be
  - need to **educate and train** a new generation of power users
- ecology
  - uncontrolled **copying** of data sets is not ecological
  - need to take care of **green computing** principles

- 
- ❑ Group and Motivation
  - ❑ The Research Data World
  - ❑ Opportunities and Challenges
  - ❑ Collaborative Data Infrastructure
  - ❑ Relevant Aspects
  - ❑ **Vision 2030**
  - ❑ Action Points

# Vision 2030

All **stakeholders**, from scientists to national authorities to general public are aware of the critical importance of preserving and sharing reliable data produced during the scientific process.

**Researchers** and practitioners from any discipline are able to find, access and process the data they need. They can be confident in their ability to use and understand data and they can evaluate the degree to which the data can be trusted.

**Producers** of data benefit from opening it to broad access and prefer to deposit their data with confidence in reliable repositories. A framework of repositories work to international standards, to ensure they are trustworthy.



# Vision 2030

**Public funding** rises, because funding bodies have confidence that their investments in research are paying back extra dividends to society, through increased use and re-use of data.

The innovative power of **industry and enterprise** is harnessed by clear and efficient arrangements for exchange of data.

The **public** has access and can make creative use of the huge amount of data available; it can also contribute to the data store and enrich it.

**Policy makers** can make decisions based on solid evidence, and can monitor the impacts of these decisions.

**Global governance** promotes international trust and interoperability.



- ❑ Group and Motivation
- ❑ The Research Data World
- ❑ Opportunities and Challenges
- ❑ Collaborative Data Infrastructure
- ❑ Relevant Aspects
- ❑ Vision 2030
- ❑ **Action Points**

# Action Points

- HLEG requests
  - need a CDI initiative
  - earmark **additional funds**
  - develop new ways to measure data **value** and reward researchers
  - train a new generation of **data scientists** and broaden understanding
  - think **green**
  - establish a **high level coordination group**
  - asking for **global collaboration**
- Recent Actions
  - **EUDAT** received grant to work on CDI
  - together with OpenAIRE work on establishing a **Data Access and Interoperability Task Force**
  - EC call for collaboration with US projects
  - DAITF preparation workshop in Copenhagen (20/21. March)
  - new e-IRG document on Data Management

- Recent EC Statements (Communication Dec 2011)
  - strong statement for **open and aggregated (meta) data**
  - EC will invest in **data infrastructures** which are based on **distributed and participatory architectures**
    - robust networks with national, regional and domain specific hubs
    - open data portals and platforms (European Data Portal)
    - supporting research & innovation for re-using, re-purposing data
  - changing the **legislation** (default = re-usage)
  - lowering **fees** to dissemination costs



Thanks for your attention.



# EUDAT

## Towards a pan-European Collaborative Data Infrastructure

Damien Lecarpentier  
CSC-IT Center for Science, Finland  
DATA2012, Indianapolis



# Outline of the talk

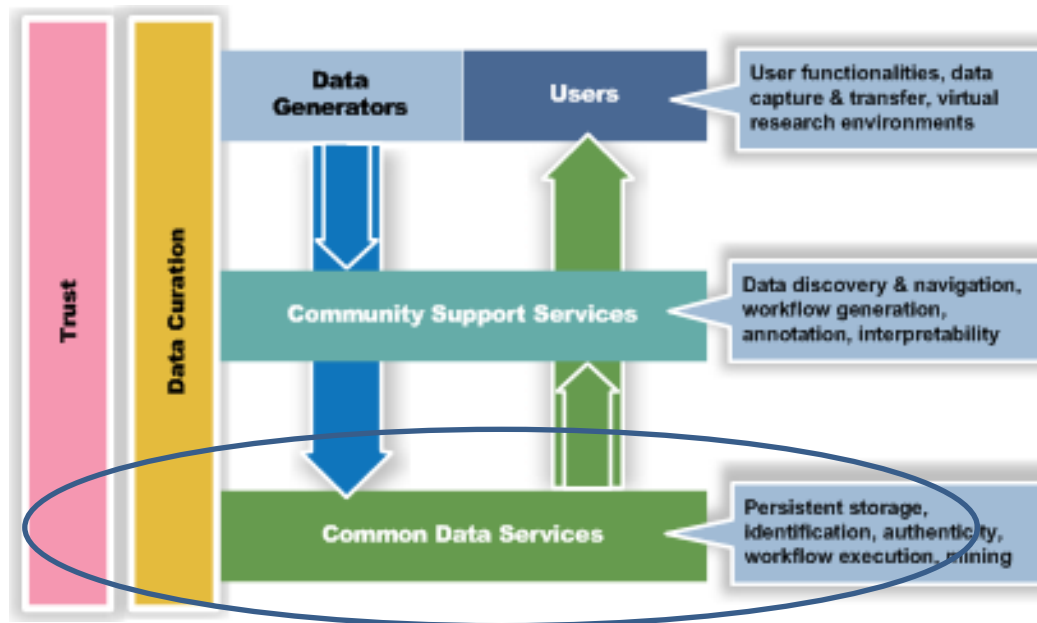
- ❑ EUDAT concept
- ❑ EUDAT consortium
- ❑ EUDAT service approach
- ❑ Some challenges ahead

# EUDAT Key facts

Project Name	EUDAT – European Data
Start date	1st October 2011
Duration	36 months
Budget	16,3 M€ (including 9,3 M€ from the EC)
EC call	Call 9 (INFRA-2011-1.2.2): Data infrastructure for e-Science (11.2010)
Participants	25 partners from 13 countries (national data enters, technology providers, research communities, and funding agencies)
Objectives	“To deliver cost-efficient and high quality Collaborative Data Infrastructure (CDI) with the capacity and capability for meeting researchers’ needs in a flexible and sustainable way, across geographical and disciplinary boundaries.”



# The CDI concept





# EUDAT Core Service Areas

## Community-oriented services

- Simple Data Access and upload
- Long term preservation
- Shared workspaces
- Execution and workflow (data mining, etc.)
- Joint metadata and data visibility

## Enabling services (making use of existing services where possible)

- Persistent identifier service (EPIC, DataCite)
- Federated AAI service
- Network Services
- Monitoring and accounting

**Core services are building blocks of EUDAT's Common Data Infrastructure**  
mainly included on bottom layer of data services



# Consortium



# Consortium





# Research Communities

ESFRI

**CLARIN**  
Common Language Resources and Technology Infrastructure



January 20, 2011  
New Virtual Language Observatory launch

Semantic data description and descriptive metadata are vital factors for determining if the data can be reused in the future. These metadata are still dependent on rapidly changing ontologies and terminologies.

**John Marks**  
ESF 2008

Activities  
Publications  
Solutions  
Laboratory  
Consultancy  
Virtual Language Observatory

Internal Web Site

**EPOS**  
EUROPEAN PLATE OBSERVING SYSTEM

Research Infrastructure and E-Science for Data and Observatories on Earthquakes, Volcanoes, Surface Dynamics and Tectonics

SEARCH



Mission & Vision   Objectives   Architecture   Partners   Preparatory Phase   Data Products

**LIFEWATCH**  
e-science and technology infrastructure for biodiversity data and observatories

Home   Contact   About   Participants   Get Involved   News   Cases   Events   Press   Documents

**LIFEWATCH COUNTRIES**  
Austria   Belgium   Denmark   Finland   France   Greece   Hungary   Italy   Netherlands   Norway   Poland   Portugal   Romania   Slovak Republic   Slovenia   Spain   Sweden   Turkey   United Kingdom

**LIFEWATCH NEWS**  
2011-02-16 **LIFEWATCH RESEARCH INFRASTRUCTURE STARTS CONSTRUCTION IN 2011** - The initial country consortium establishing the Lifewatch research infrastructure agreed to finance ... [Read more](#)  
2011-01-19 **LIFEWATCH CLOSING EVENT** - On this page you can download all the slides presented at the closing event of the Lifewatch preparatory project a first group of ... [Read more](#)  
2011-01-17 **LIFEWATCH CONSTRUCTION KICKS OFF ON JANUARY 19TH** - On 19 January 2011, at the closing conference of the Lifewatch preparatory project a first group of ... [Read more](#)

**LIFEWATCH FOCUS**  
**Lifewatch research infrastructure starts construction in 2011**  
The initial country consortium establishing the Lifewatch research infrastructure agreed to finance the start-up activities for the infrastructure construction. These countries will host the Common Facilities of Lifewatch.  
On 19th January 2011 representatives from organizations in Hungary, Italy, the Netherlands, Romania and Spain signed a Memorandum of Understanding to cooperate for an early start of the Lifewatch infrastructure for biodiversity and ecosystem research. The Lifewatch Stakeholders Board, representing the ten countries aiming at establishing the Lifewatch ERIC, welcomed the initiative to start early construction.

Newsletter  
Subscribe to our newsletter. Send an email to [newsletter@lifewatch.eu](mailto:newsletter@lifewatch.eu)

Quote  
"Through our Memorandum of Cooperation GBIF and Lifewatch based on our respective complementary mandates, now have a formal framework for co-operation and collaboration on infrastructural developments, building on GBIF's 10 years of investment to date."  
**Dr. Nick King**  
Director Global Biodiversity Information Facility (GBIF)

**enes**  
European Network for Earth System Modelling

Welcome

News

IS-ENES

The Rationale

The Aims

The MOU

Related Projects

Partners

Meetings

Sitemap

Contact

Imprint

search...

ENES > Welcome

print page

Welcome

**ENES Townhall Meeting at EGU 2010: Here is the announcement!**

For latest news on IS-ENES click [here!](#)

A major challenge for the climate research community is the development of comprehensive Earth system models capable of simulating natural climate variability and human-induced climate changes. Such models need to account for detailed processes occurring in the atmosphere, the ocean and on the continents including physical, chemical and biological processes on a variety of spatial and temporal scales. They have also to capture complex nonlinear interactions between the different components of the Earth system and assess, how these interactions can be perturbed as a result of human activities.

Accurate scientific information is required by government and industry to make appropriate decisions regarding our global environment, with direct consequences on the economy and lifestyles. It is therefore the responsibility of the scientific community to accelerate progress towards a better understanding of the processes governing the Earth system and towards the development of an improved predictive capability. An important task is to develop an advanced software and hardware environment in Europe, under which the most advanced high resolution climate models can be developed, improved, and integrated.

**Virtual Physiological Human**  
network of excellence

Home   WP1   WP2   WP3   WP4   WP5   VPH-I   MIP   Login

Search: ...

Building a wider VPH community

**HIGHLIGHTS**  
Interface Focus special issue with best papers from the VPH2010 Conference  
VPH-NoE and the Pistoia Alliance  
Exemplar Project Call 3!  
Join the Public Forum of the VPH-FET Support Action  
Multi-institutional Graduate Programme for Virtual Physiological Human Scientists (VPH-MGP)  
VPH Vision & Strategy Paper I  
VPH NoE 2010 Newsletter (Jan 2010) now available

**LATEST VPH EVENTS**  
01.06.2011 - 03.06.2011 ICOS 2011 (Toulon)  
06.06.2011 - 09.06.2011 VPH08  
08.06.2011 - 10.06.2011 VPH08  
08.06.2011 - 10.06.2011 VPH08

The VPH NoE is a project which aims to help support and progress European research in biomedical modelling and simulation of the human body. This will improve our ability to predict, diagnose and treat disease, and have a dramatic impact on the future of healthcare, the pharmaceutical and medical device industries.

**VPH 2010**  
September 30th - October 1st 2010  
Brussels, Belgium

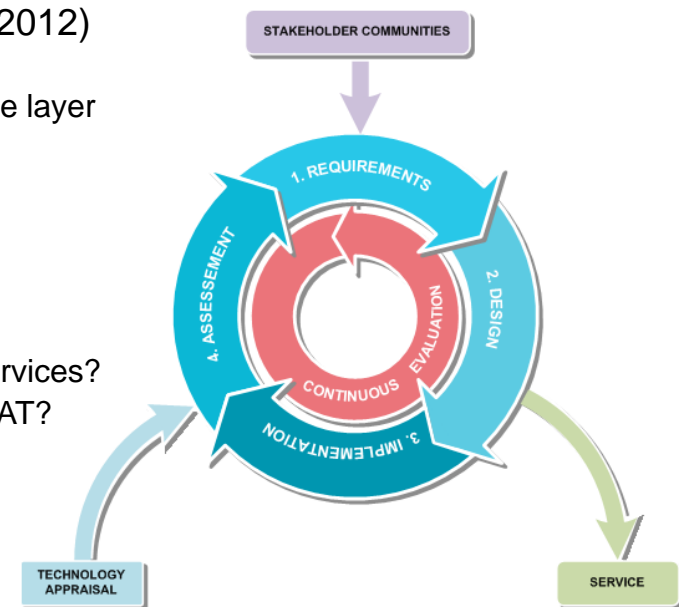
# EUDAT service design activities

## 1. Capturing Communities Requirements (WP4)

- 1st round of interviews with the five initial communities (Oct.-Dec. 2012)
  - Understand how data is organised in each community
  - Collect first wishes and specific requirements from a common data service layer
- Next phase: refine analysis and expanding it to other communities

## 2. Building the corresponding services (WP5)

- Technology appraisal (ongoing)
  - What is already available at partners's sites to build the corresponding services?
  - What are the gaps and market failures that should be addressed by EUDAT?
- Next phase: Developing candidate services
  - Adapt services to match the requirements
  - Integrate with community and SP services
  - Test and evaluate with communities



## 3. Deploying the services and operating the federated infrastructure (WP6)

- Designing the federated infrastructure and the interfaces for cross-site operations (ongoing)
- Next phase: integrating and coordinating resource provision, operations and support

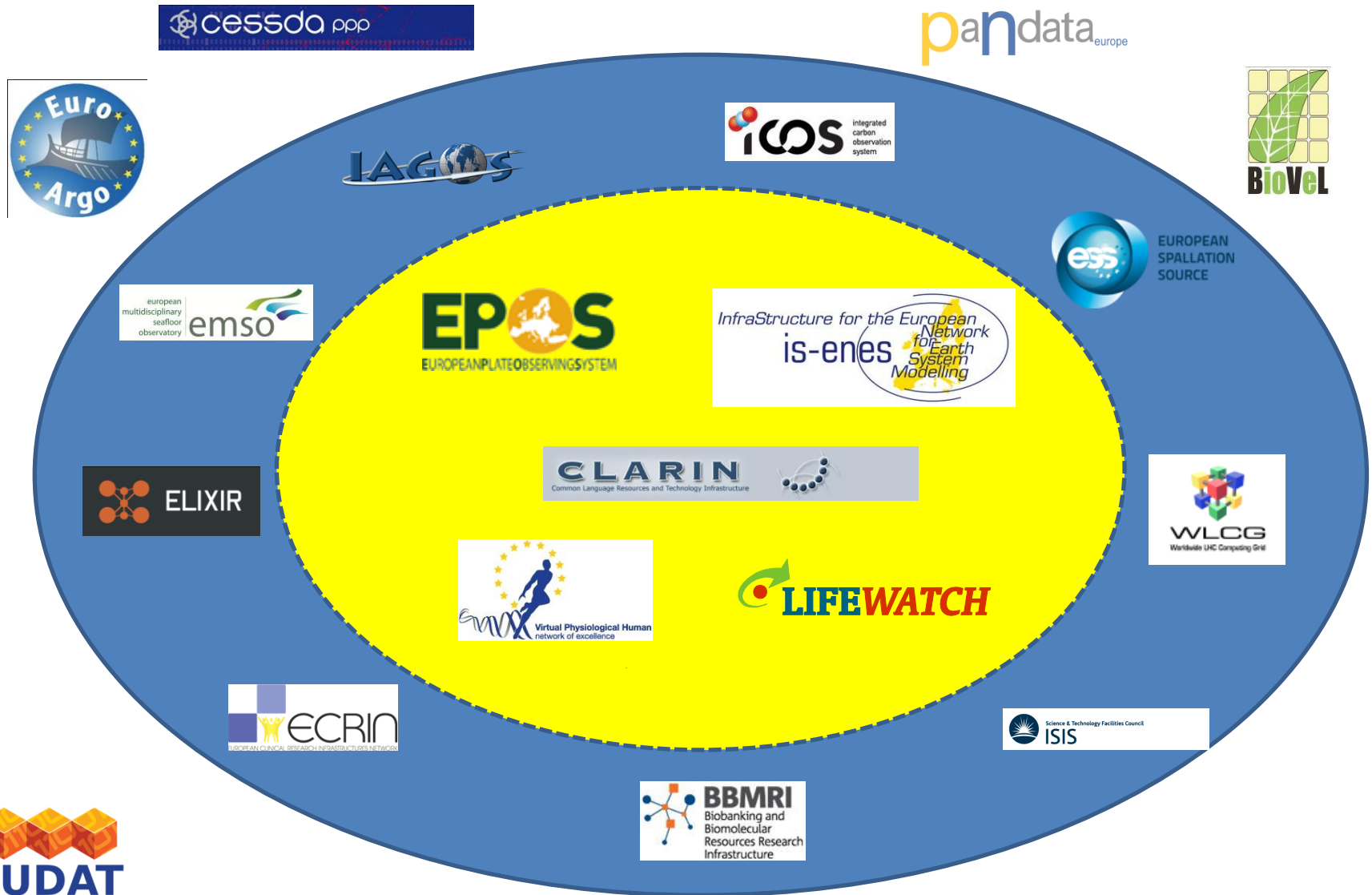


# First Service Cases

## ■ November 2011: shortlist of 6 service/use cases identified

- **Safe replication:** Allow communities to replicate data to selected data centers for storage and do this in a robust, highly available way. **Under implementation**
- **Dynamic replication:** Perform (HPC) computations on replicated data. Move (part of) the safely replicated data to powerful machines and move the results back into the archives. **Under implementation**
- **AAI:** A solution for a working AAI system in a federation scenario.
- **Metadata:** A joint metadata domain for all data that is stored by EUDAT data centers by harvesting metadata records for all data objects from the communities. Allow to have a catalogue to demonstrate what EUDAT stores, and to have a registry which can be used for automatic operations such as data mining.
- **PID:** a robust, highly available and effective PID system that can be used within the communities and by EUDAT.
- **Research data store:** A function that will help researchers mediated by the participating communities to upload and store data which is not part of the officially handled data sets of the community.

# Reaching out to other communities



# Research fields

<b>Environmental Science</b>	<b>ENES, EPOS, Lifewatch, EMSO, IAGOS-ERI, ICOS, Euro-Argo</b>
<b>Social Sciences and Humanities</b>	<b>CLARIN, CESSDA, DARIAH</b>
<b>Biological and Medical Science</b>	<b>VPH, ELIXIR, BBRMI, ECRIN, DiXA</b>
<b>Physical Sciences and Engineering</b>	<b>WLCG, ISIS, PanData</b>
<b>Material Science</b>	<b>ESS...</b>

EUDAT targets all scientific disciplines (discipline neutral):

- To enable the capture and identify cross-discipline requirements
- To involving the scientists of all the communities in the shaping of the infrastructure and its services



# Some forthcoming challenges

## ■ Scaling

- Can we expect that the requirements identified will be shared by other research communities?
- How to maintain a high level of interoperability in the context of diversity of data, disciplines and practices?
- How far shall the geographical pan-European dimension be sought for?

## ■ Collaboration models

- What kind of collaboration and partnership can we offer to interested stakeholders?
- How will the future infrastructure work with and interact with other infrastructures and projects?
- How to move for a project collaboration to a federated infrastructure?

## ■ Funding models

- How will the infrastructure be funded in the future?
- How far are the public bodies (EC and MS) willing to fund the infrastructure?
- Shall users pay for the service they are provided, and if so on what basis (pay per use, membership fee, etc.)?

# EUDAT User Forum



**EUDAT - 1<sup>ST</sup> USER FORUM**  
TOWARDS A COLLABORATIVE DATA INFRASTRUCTURE  
INVESTIGATING RESEARCH COMMUNITIES REQUIREMENTS

**Place:**  
**Jordi Girona Street, No.31**  
**Rectorat Building, Junes room,**  
**08034, Barcelona**

## DAY 1: 7TH MARCH 2012

12.00	Welcome and Snacks	
13.00	<b>SESSION 1 - SETTING THE SCENE</b>	<b>CHAIR: KIMMO KOSKI</b>
13.00	EUDAT - an Overview for a User Perspective	Damien Lecarpentier
13.15	CDI and EUDAT	Peter Wittenburg
13.30	Climate modeling and EUDAT	Michael Lautenschlager
13.45	Seismology and EUDAT	Alberto Michelini
14.00	Physiology and EUDAT	Stefan Zasada
14.15	Linguistics and EUDAT	Pavel Straňák/Daan Broeder
14.30	Discussion	
15.00	Coffee Break	
15.30	<b>SESSION 2 - EUDAT SERVICE CASES</b>	<b>CHAIR: ALBERTO MICHELINI</b>
15.30	Replication service and its requirements	Peter Wittenburg
15.40	Staging Replicas for computations and requirements	Stefan Zasada
15.50	Researchers' Data Store and requirements	Daan Broeder
16.00	Joint EUDAT Metadata Domain	Michael Lautenschlager
16.10	The Risky Tasks in EUDAT	David Corney
16.20	Discussion	
16.45	Short break	
17.00	Community Presentations I	
18.30	End	
20.30	Dinner	

## DAY2: 8TH MARCH 2012

09.00	<b>SESSION 3 - ENABLING TECHNOLOGIES</b>	<b>CHAIR: MICHAEL LAUTENSCHLAGER</b>
09.00	Distributed Authentication - will it work	Mark van de Sanden
09.10	PID Systems for Digital Objects	Ulrich Schwardmann
09.20	Replication Technologies	John Kennedy
09.30	Hosting Services and Staging Data	Johannes Reetz
09.40	Creating a Joint Semantic Domain	Peter Wittenburg
09.50	Other Technology and Operation Issues	M. van de Sanden/J. Reetz
10.00	Discussion	
10.30	Coffee Break	
11.00	<b>SESSION 4 - EUDAT AND THE WAY FORWARD</b>	<b>CHAIR: PAVEL STRAŇÁK</b>
11.15	Sustaining the infrastructure	Alison Kennedy
11.25	Training the new data scientist	Nagham Salman
11.35	Discussion	
12.00	Community Presentations II	
12.30	<b>WRAP UP AND GENERAL DISCUSSION</b>	<b>K. KOSKI/ P. WITTENBURG</b>
13.00	End	



# Welcome to the 1st EUDAT Conference



5-8 November 2012, Barcelona

- A high level international event where EUDAT first results will be demonstrated
- A forum to discuss the future of EUDAT and data infrastructures
- 2nd EUDAT User Forum
- Training tutorials

# EU-US collaboration

## ■ DAITF

- Data Access and Interoperability Task Force: to create a global interaction framework, pushing harmonization and standardisation with respect to the abstract data architecture and all its essential components.
- Data architecture workshop in Copenhagen (ICRI conference)

## ■ iCORDI (new project proposal)

- New 2M€ proposal submitted to the EC in November 2012 (results over eligibility for funding tba shortly)
- Project goal: to establish a **coordination platform** between Europe and the USA to discuss and improve the interoperability of today's and tomorrow's scientific data infrastructures of both continents.
  - fostering discussion between relevant stakeholders in the EU and US over concrete topics related to the interoperability of the data architectures and solutions based on a top-down approach;
  - overcoming the identified challenges and turning the areas of convergence into concrete specifications that can be immediately implemented on both continents by bringing data practitioners together in a bottom-up process;
  - demonstrating through concrete examples of collaboration what works and what are the remaining barriers and challenges to be tackled to achieve full interoperability.

# EUDAT management team



Damien Lecarpentier, CSC,  
Project Manager



Kimmo Koski, CSC  
Project Coordinator



Peter Wittenburg, MPI-PL  
Scientific Coordinator  
Stakeholders requirements



Nagham Salman, BSC,  
Dissemination



Alison Kennedy, EPCC  
Sustainability



Mark van de Sanden, SARA  
Services



Johannes Reetz, RZG  
Operations



David Corney, STFC  
Scalability



# **The Board on Research Data and Information (BRDI)**

**NSF Principal Investigators Meeting for DataNet and INTEROP  
Indianapolis, IN  
January 26, 2012**

Prepared by:

Paul F. Uhler, Director, BRDI, [puhler@nas.edu](mailto:puhler@nas.edu)

Presented by:

Daniel Cohen, Program Officer, BRDI (on detail from the Library of Congress), [dcohen@nas.edu](mailto:dcohen@nas.edu)

**Projects funded by NSF grant OCI 1040898**

# Board on Research Data and Information

The Board's *mission* is to improve the stewardship, policy, and use of digital data and information for science and the broader society.

Formed in 2008, the Board interacts broadly with federal agency sponsors and the various stakeholders in the research community to make progress on the research data and information priorities that are essential for our nation's future. The Board also represents to the US National Committee for CODATA, an interdisciplinary committee of ICSU.

The Board is co-chaired by Fran Berman, RPI, and Clifford Lynch, Coalition for Networked Information. It has 20 senior members from academia and industry, and 5 staff.

Federal science and informatics agency sponsors currently include: NSF, NIH, NIST, NOAA, IMLS, USGS, and the Library of Congress.



# Board on Research Data and Information

**BRDI undertakes the following tasks within its primary mission areas:**

- Addresses emerging issues in the management, policy, and use of research data and information at the national and international levels.
- Through studies and reports of the National Research Council, provides independent and objective advice, reviews of programs, and assessment of priorities concerning research data and information activities for its sponsors.
- Encourages and facilitates collaboration across disciplines, sectors, and nations with regard to common interests in research data and information activities.
- Monitors, assesses, and contributes to the development of U.S. government and research community positions on research data and information programs and policies.
- Initiates or responds to requests for consensus studies, workshops, conferences, and other activities within the Board's mission, and provides oversight for the activities performed under the Board's auspices.
- Broadly disseminates and communicates the results of the Board's activities to its stakeholders and to the general public.



# Board on Research Data and Information

## Recently completed projects:

- *U.S.-China Roundtable on Scientific Data Cooperation* (2006-2011)
- *Symposium on Common Use Licensing of Publicly Funded Scientific Data and Publications* (held in Beijing and Taipei, March 2009)
- *International Workshop on the Socioeconomic Effects of Public Sector Information on Digital Networks: Toward a Better Understanding of Different Access and Reuse Policies* (report published in June 2009)
- *Symposium on the Data Sharing Plans for GEOSS and on the Benefits of Data Sharing for Science* (with the Group on Earth Observations, November 2009)
- *International Symposium on Designing the Microbial Research Commons: New Strategies for Accessing, Managing, and Using Essential Public Knowledge Assets* (report published in September 2011)



# Board on Research Data and Information

## Ongoing projects:

- *Board meetings and public mini-symposia on select topics* (ongoing -- twice per year)
- *U.S. Committee for CODATA* (ongoing -- represents the US interests to the international CODATA on interdisciplinary scientific data issues, including participation in conferences and Task Groups)
- *Forum on CODATA-World Data System Cooperation* (meeting series)
- *BRDI Sponsor Forum* (meeting series)
- *National Symposium and Workshop on the Future of Scientific Knowledge Discovery in Open Networked Environments* (held March 2011, report due spring 2012)
- *International Symposium on The Case for International Sharing of Scientific Data—A Focus on Developing Countries* (held April 2011, report due 2012)
- *Developing Data Attribution and Citation Practices and Standards* (2011-2013, with two reports)
- *Future Career Opportunities and Educational Requirements for Digital Curation—A Consensus Study* (2011-2013)

# Board on Research Data and Information

## Planned projects, 2012-2013

- *Sustainability Strategies for Publicly Funded Research Databases—A Consensus Study*
- *Research Data as Intellectual Property: How the Law Influences Data Sharing—A National Symposium and Workshop*
- *Global Research-data Access Interoperability and Policies* (with an EU group, if approved)
- *The Value and Impact of Public Scientific Data*

See [www.nas.edu/brdi](http://www.nas.edu/brdi) for more information about the projects funded by NSF OCI.



# Geospatial Semantic Interoperability

---

*INTEROP – Spatial Ontology Community of Practice:  
an Interdisciplinary Network to Support  
Geospatial Data Sharing, Integration, and Interoperability*

Nancy Wiegand  
*University of Wisconsin - Madison*



**SOCOP INTEROP  
CYBERINFRASTRUCTURE**



NSF Meeting January 2012



# SOCoP INTEROP

- Goals
  - Apply and develop semantic technologies for the Geospatial domain
  - Share ontologies to promote data interoperability
- Submitted by 8 members of SOCoP :
  - **Gary Berg-Cross** - Knowledge Strategies/SOCoP officer
  - **Mike Dean and Dave Kolas** – Raytheon BBN Technologies
  - **John Moeller** - JJMoeller and Associates/SOCoP officer
  - **Nancy Wiegand** - University of Wisconsin-Madison,
  - **James Wilson** - James Madison University
  - **Peter Yim** - CIM3 Engineering, Inc.
  - **Naijun Zhou** - University of Maryland College Park



# Spatial Ontology Community of Practice

---

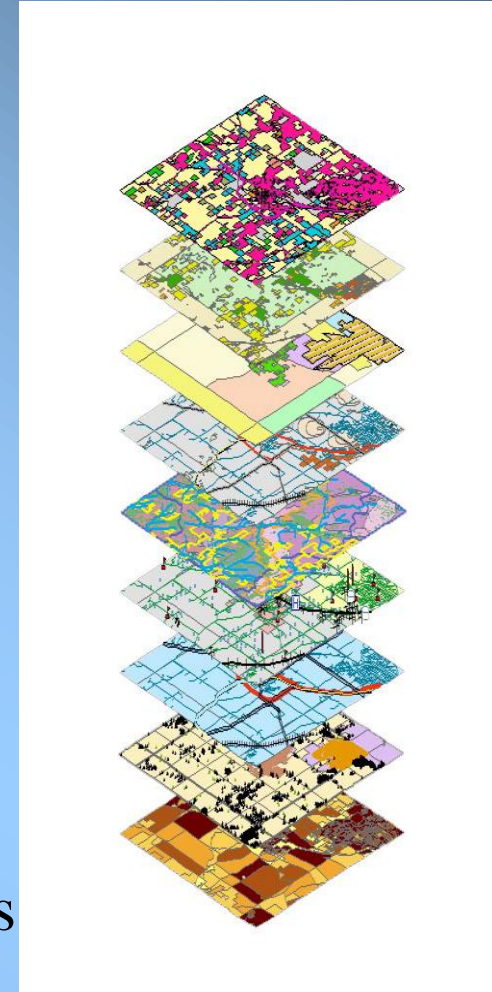
- **SOCoP** – National level group of practitioners, academic researchers, federal agency workers, industry representatives
- Started in 2006, <http://www.socop.org>
- Recognize the need for **semantic** interoperability for geospatial data and the potential of **knowledge bases** and formal representations to help solve semantic heterogeneity
- Monthly conference calls, usually the 3<sup>rd</sup> Wed.



**SOCOP INTEROP  
CYBERINFRASTRUCTURE**

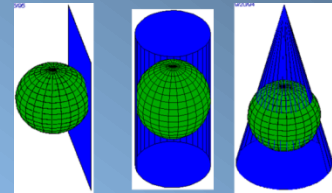
# Overlay – Integrate Layers

- **Basic reference** - geodetic, topographic, “base map” ...
- **Cadastral** - land ownership, tax assessment, land tenure...
- **Administrative** - jurisdictions, zones, tracts...
- **Resource** - soils, land cover, land use, hydrography...
- **Infrastructure** - transportation, sewer, water, electric, cable...
- **Imagery** - aerial photography, satellite images, lidar...



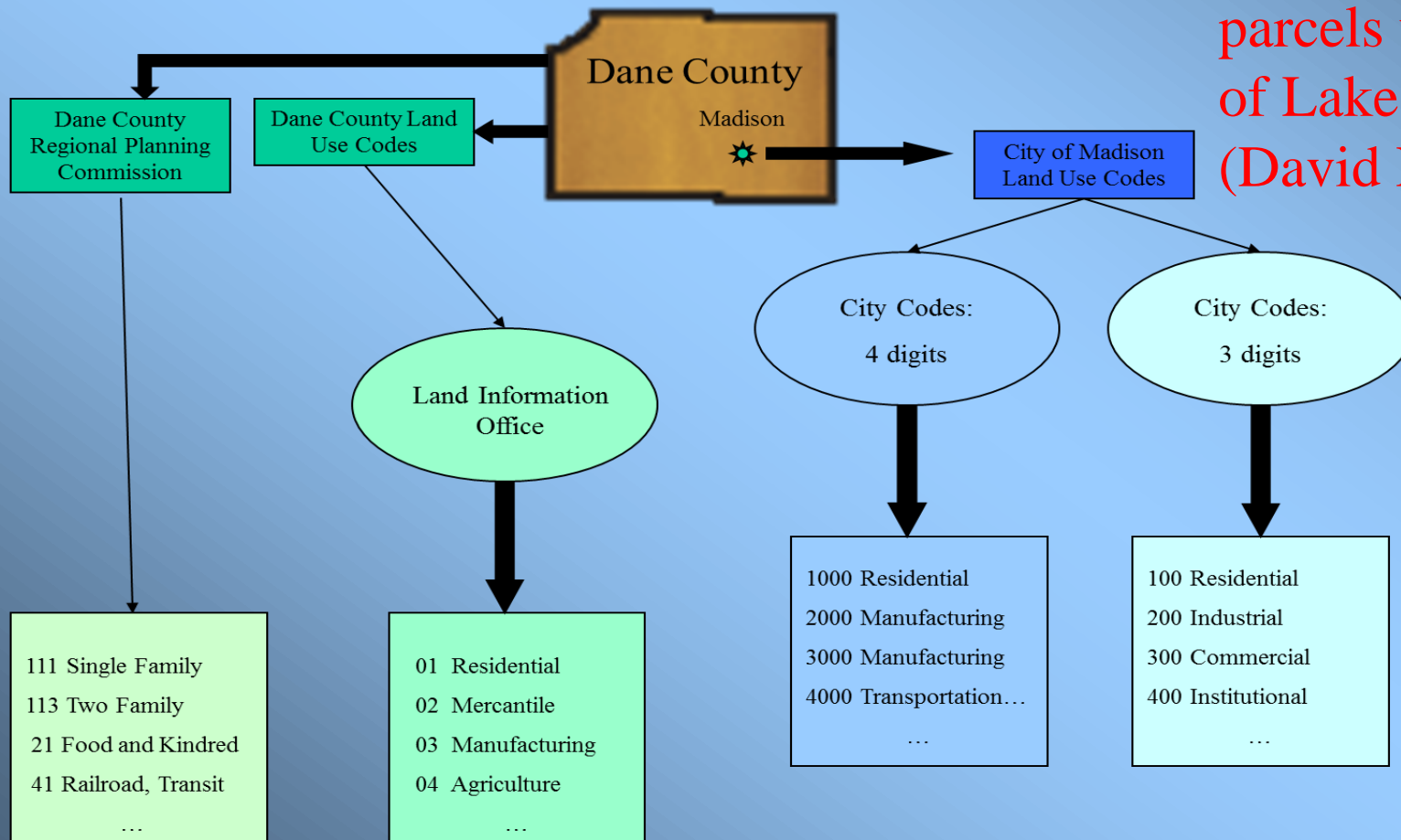
# Integration Problems

- Integrating by layers - data need to be in the same projection. This problem is mostly already solved; software will automatically re-project data.
- But, integrating geospatial data still has **semantic problems**, e.g.,
  - **horizontal** geographic integration (e.g., each county has a different land use coding system)
  - integration **through time** (e.g., categories differed in the past)
- Need automatic ‘**semantic re-projections**’



# Semantic Heterogeneity in Land Use Data

**Example High Level Categories in Land Use Codes.**



Find and aggregate  
types of land use in  
parcels within 1000'  
of Lake Michigan  
(David Hart).

-Local data  
Providers

-Classifica-  
tions change  
over time.



# Vision

---

- Improve **querying** in geospatial data
  - Ontologies for semantic interoperability
- Improve **search** for geospatial data and services
  - Ontologies for metadata and background knowledge
- Semantic components have a role in **geo-architectures**
- Geospatial data in the **Linked Open Data** cloud



# INTEROP Tasks - Overview

- Cyberinfrastructure
  - Web presence – Wiki ([www.socop.org](http://www.socop.org))
  - Open Ontology Repository (OOR)
  - GeoSPARQL
  - Educational component
- Other
  - Prototypes/demos
  - Workshops/meetings



Open Ontology Repository

Browse Search Mappings Recommender Projects

Recently Viewed | Index | Help | Feedback

**Browse**

Access all ontologies that are available in SOCOP OOR. You can filter this list by category to display ontologies relevant for a certain domain. You can also filter ontologies that belong to a certain group. Subscribe to the SOCOP OOR RSS feed to receive alerts for submissions of new ontologies, new notes, and new projects. You can subscribe to feeds for a specific ontology at the individual ontology page. Add a new ontology to SOCOP OOR using the Submit New Ontology link.

FILTER BY CATEGORY: All Categories

FILTER BY GROUP: All Groups

FILTER BY TEXT:

ONTOLOGY NAME	VISIBILITY	TERMS	NOTES	REVIEWS	PROJECTS	UPLOADED	AUTHOR
Basic Formal Ontology (BFO)	Public	26	0	0	0	12/02/2011	Holger Steneshorn
DOLCE-DOLce Ultimate (DUA)	Public	25	0	0	0	12/02/2011	Aldo Gangemi
Friend of a Friend (FOAF)	Public	3	0	0	0	09/19/2011	Dan Brickley
GeoNames (GeoNames)	Public	10	0	0	0	01/15/2012	Bernard Vatant
GeoPARQL (GeoPARQL)	Public	0	0	0	0	09/28/2011	Matthew Perry
ISO 19103 Geographic Information Conceptual Schema Language (ISO19103)	Public	43	0	0	0	12/02/2011	Bora Beran, Akim Saifal Islam, Luis Bermudez, Stephane Fellah, Michael Piascecki
ISO 19107-2003 Geographic Information - Spatial Schema (ISO19107)	Public	102	0	0	0	12/02/2011	Bora Beran, Akim Saifal Islam, Volkan Yargic, Stephane Fellah, Michael Piascecki
ISO 19108-2003 Geographic Information - Temporal Schema (ISO19108)	Public	43	0	0	0	01/15/2012	Akim Saifal Islam, Bora Beran, Luis Bermudez, Stephane Fellah, Michael Piascecki
ISO 19111-2003 Geographic Information - Spatial Referencing by Coordinate (ISO19111)	Public	12	0	0	0	01/15/2012	Akim Saifal Islam, Bora Beran, Michael Piascecki
ISO 19112-2003 Geographic Information - Spatial Referencing by Geographic Identifier (ISO19112)	Public	7	0	0	0	01/15/2012	Akim Saifal Islam, Bora Beran, Michael Piascecki
ISO 19115-2003 Geographic Information - Metadata (ISO19115)	Public	120	0	0	0	01/15/2012	Akim Saifal Islam, Luis Bermudez, Bora Beran, Stephane Fellah, Michael Piascecki
ISO 19115-2003 Geographic Information - Metadata Application (ISO19115-APP)	Public	9	0	0	0	01/15/2012	Akim Saifal Islam, Michael Piascecki
ISO/TS 19109 Geographic Information - Rules for Application Schema (ISO19109)	Public	10	0	0	0	01/15/2012	Akim Saifal Islam, Bora Beran, Michael Piascecki
ISO/TS 19110 Geographic Information - Methodology for Feature Cataloguing (ISO19110)	Public	11	0	0	0	01/15/2012	Akim Saifal Islam, Bora Beran, Michael Piascecki
Land Use and Cover Change (LUCC)	Public	216	0	0	0	09/19/2011	Gary Berg-Cross, Dawn Parker, Gary Pothil
NeoGeo Geometry Ontology (NeoGeo-Geometry)	Public	10	0	0	0	01/15/2012	Juan Martin Salas, Andreas Harth, Barry Norton, Luis M. Vlachos, Alexander De Leert, John Goodwin, Claus Stadler, Sachin Anand, Dominic Harries
							Juan Martin Salas, Andreas Harth, Barry

# Workshops We Organized

---

- **Spatial Semantics and Ontologies (SSO)**  
Workshop at ACM SIGSPATIAL GIS 2011  
Conference, Nov. 1, 2011
- **Terra Cognita** Workshop at the International  
Semantic Web Conference, October 2011
- **SOCoP** Workshop at the USGS in Reston, VA,  
Dec. 2, 2011
- **GeoVoCamp**, Washington, D.C., June 2011
- **SOCoP** Workshop Dec. 3, 2010, D.C. area



Thank you!



The SOCoP INTEROP Team



# TERRA POPULUS

A Global Population/Environment Data Network





# The explosion of Scientific Data

*Because of the massive decline in the cost of data collection, storage, and analysis, the quantity of scientific data being collected is growing at an extraordinary pace*

- New opportunities for analysis
- New methods are being applied
- Marked acceleration in the pace of discovery



# The Big Challenges

*The quantity of scientific data is exploding, but we lack basic infrastructure to maintain them or capitalize on opportunities for analysis and discovery*

- Most scientific data is at risk of loss
- Most scientific data is inaccessible
- Metadata are usually incomplete and inadequate
- Little interoperability across datasets or data types
- Data are trapped in disciplinary silos



# TerraPop Goals

Provide an organizational and technical framework to preserve, integrate, and disseminate global-scale spatiotemporal data describing population and the environment.

- Census microdata
- Government land-use statistics
- Land cover data from satellite imagery
- Historical climate records (temperature, precipitation, cloud cover)



Age  
Sex  
Relationship  
Race  
Birthplace  
Mother's birthplace  
Occupation

H9100002400000000088001001000220100
P910000020101032120010010010011504
P910000010201036220010010010011999
P910201000301011220060010010011999
P910201000301009120060010010011999
P910201000301007120060010010011999
P910201000301006120060010010011999
P910201000301004220060010010011999
P910201000301003220060010010011999
P910201000301002220060010010011999
H9100002400000000088001001000110100
P910000020101030110010290510511310
P910000010201021210010290290171999
P910201000301001110060010290291999
H9100002400000000088001001000220100
P910000020101045120010010010011100
P910000010201025220010010010011820
P910201000301007220060010010011999
H9100002400000000088001001000220100
P910000020101049120010010010011100
P910000010201049220010010010011820
P910201000301019220060010010011820
P910201000301015220060010010012820

## Microdata Structure

Geographic and housing characteristics

Household record (shaded) followed by a person record for each member of the household

For each type of record, columns correspond to specific variables



# The Power of Microdata

- **Customized measures:** Variables based on combined characteristics of family and household members, capitalizing on the hierarchical structure of the data
- **Multivariate analysis:** Analyze many individual, household, and community characteristics simultaneously
- **Interoperability:** Harmonize data across time and space

For each person, detailed information about geographic location, economic activities, educational attainment, literacy, fertility history, child mortality, migration, place of former residence, marital status, consensual unions, family composition, disabilities, water supply, sewage, building materials (floor, roof, etc.), and many other characteristics.







[Home](#) | [Select Data](#) | [FAQ](#) | [Contact](#) | [Login](#)

## PROJECT

About IPUMS-I  
How to Cite IPUMS-I  
User Registration and Login

## DATA

Browse and Select Data  
Download Your Data Extract  
GIS and Other Data Files

## SAMPLES

Sample Descriptions  
Variance Estimation  
Source Documents

## RESOURCES

International Partners  
World Data Inventory  
Microdata Handbook  
Bibliography

# Integrated Public Use Microdata Series, International

census microdata for social and economic research

IPUMS-International is a project dedicated to collecting and distributing census data from around the world. Its goals are to:

- Collect and preserve data and documentation
- Harmonize data
- Disseminate the data absolutely free!

62 countries - 185 censuses - 397 million person records

## IPUMSI News

June 2011 data release  
2010 award winners  
Improved web interface  
IPUMS Havana workshop  
June 2010 data release  
Mortality and fertility data  
NIH extends IPUMS-I  
... All news items

## MPC Data Projects

[IPUMS-USA and others](#)





# NAPP

## North Atlantic Population Project

[Home](#) [Select Data](#) [FAQ](#) [Contact](#) [Login](#) [Data Cart](#)

### PROJECT

[About NAPP](#)  
[Data Releases](#)  
[Revision History](#)  
[User Registration & Login](#)

### DATA

[Browse and Select Data](#)  
[Download Data Extract](#)  
[Linked Samples](#)  
[Online Data Analysis](#)

### DOCUMENTATION

[Samples](#)  
[Census Questionnaires](#)  
[Other Documentation](#)

### RESOURCES

[NAPP Participants](#)  
[Citation and Use](#)  
[Bibliography](#)

## North Atlantic Population Project

Census microdata from Canada, Great Britain, Germany, Iceland, Norway, Sweden, and the United States from 1801 to 1910. The project's goals are to:

- Harmonize data, including many complete count datasets
- Link individuals between census years for longitudinal analysis
- Disseminate the data absolutely free!



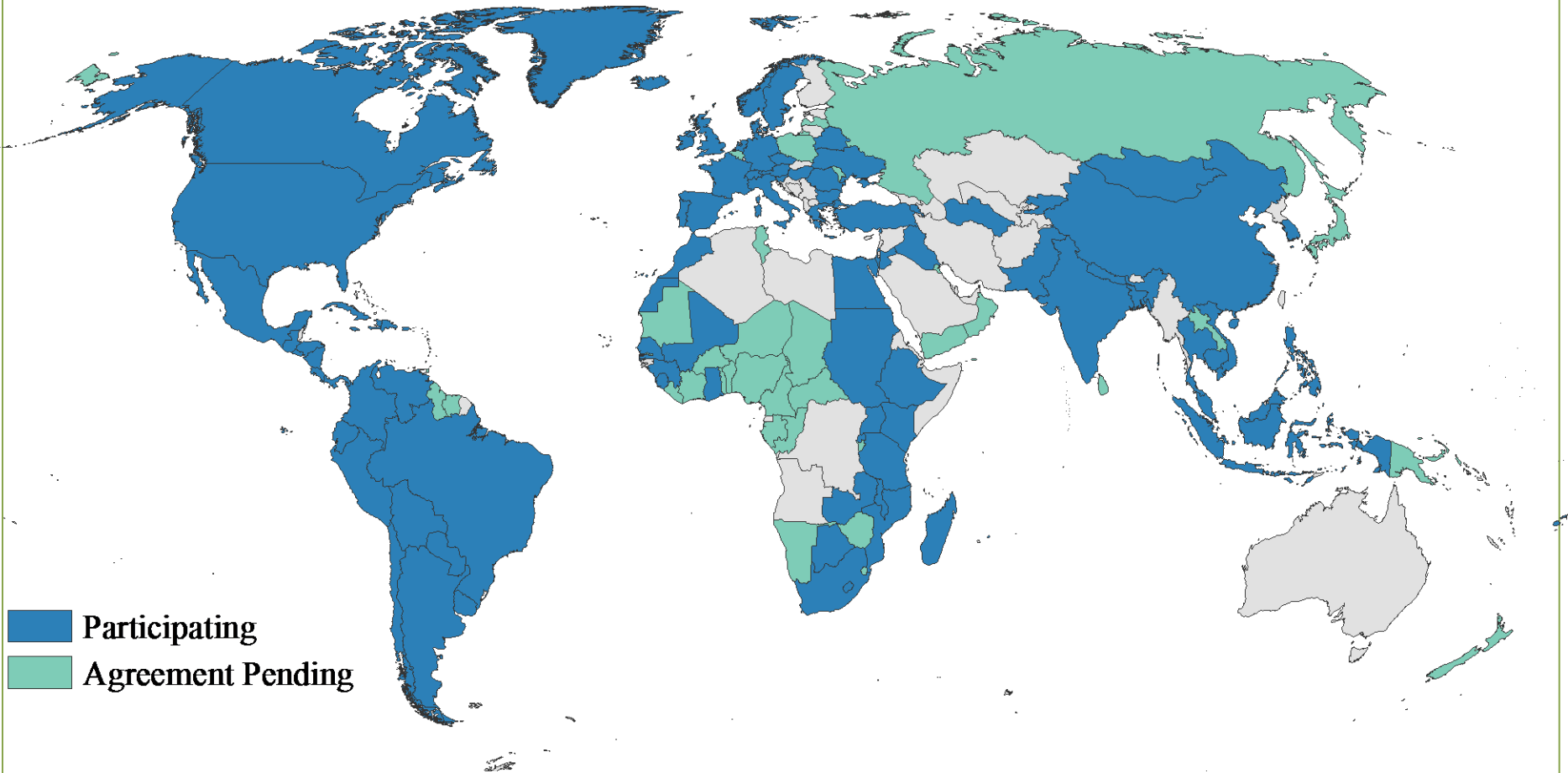
### NAPP News

**NEW!** Full-count data for Iceland 1801 and 1901  
**NEW!** Full-count data for Norway 1801  
**NEW!** Samples for Canada 1852 and 1891  
[Linked data for Norway and the U.S.](#)  
[Mecklenburg-Schwerin 1819 sample](#)

### Other MPC Projects

[IPUMS-International](#)  
[IPUMS-USA](#)  
[IPUMS-CPS](#)  
[IHIS](#)  
[NHGIS](#)

# IPUMS/NAPP Participating Countries



# IPUMS/NAPP Data Releases

2011:

- 600 Million persons
- 300 censuses and surveys
- 65 countries

2016:

- 1.2 billion persons
- 800 censuses and surveys
- 110 countries



# TerraPop Partners



INSTITUTE ON THE  
ENVIRONMENT

UNIVERSITY OF MINNESOTA  
Driven to Discover<sup>SM</sup>

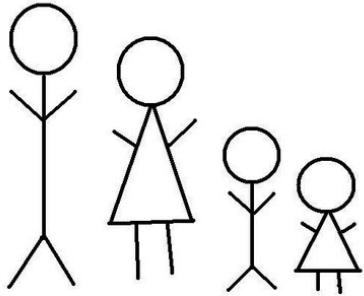


ICPSR | INTER-UNIVERSITY  
CONSORTIUM FOR  
POLITICAL AND  
SOCIAL RESEARCH



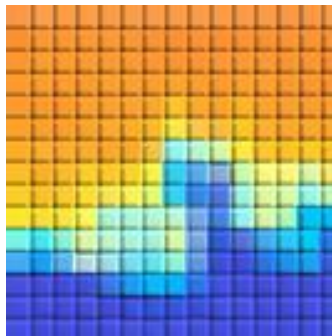
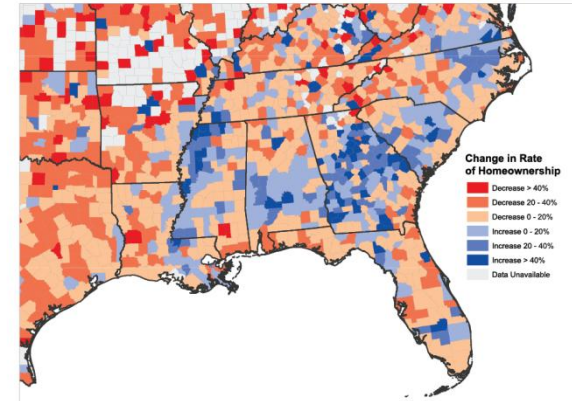


# Three data types



**Microdata:**  
Characteristics of individuals  
and households

**Small-area data:**  
Characteristics of places defined  
by administrative boundaries



**Gridded data:**  
Values arranged in rows  
and columns



# Three output formats

1. Census microdata with attached characteristics describing land use, land cover, and climate for local areas
2. Aggregate data for administrative districts with tabulated population data and environmental characteristics
3. Gridded data with characteristics of population and environment

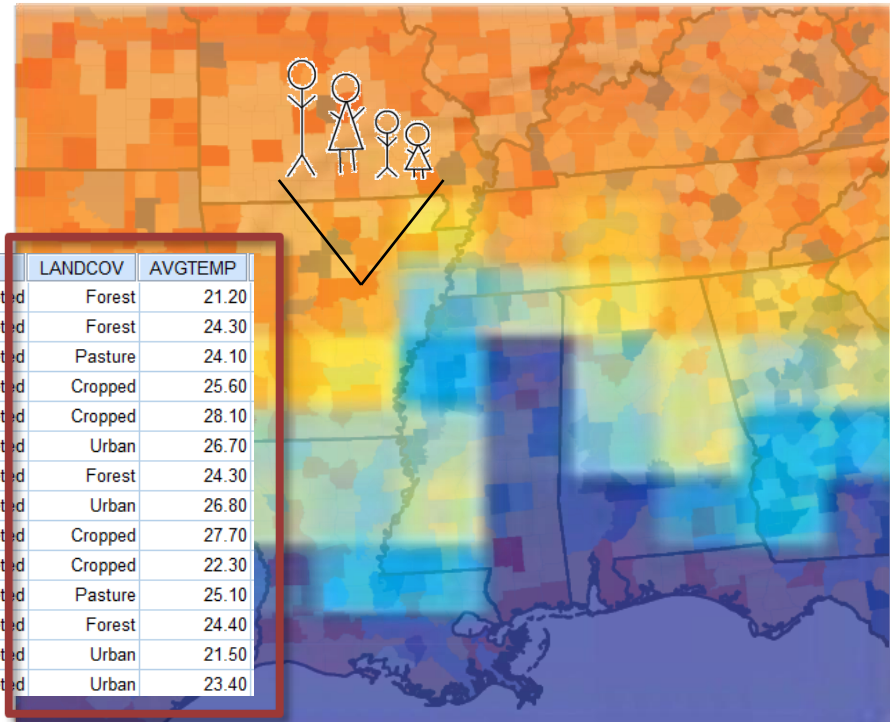


# Data Fusion – Microdata Output

Census microdata with attached characteristics describing land use, land cover, and climate for local areas

Individuals and households  
with their environmental  
and social context

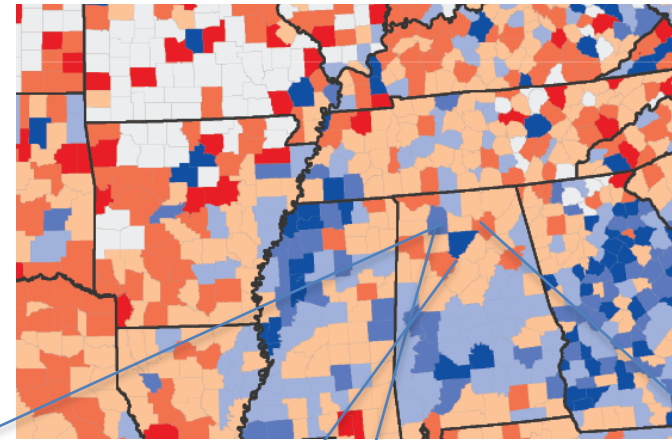
YEAR	AGE	SEX	MGRATE5	EDATTAN	LANDCOV	AVGTEMP
1991	10	Male	Same major, same minor administrative unit	Less than primary completed	Forest	21.20
1991	27	Female	Same major, different minor administrative unit	Secondary completed	Forest	24.30
1991	54	Female	Same major, same minor administrative unit	Primary completed	Pasture	24.10
1991	37	Male	Same major, same minor administrative unit	University completed	Cropped	25.60
1991	37	Female	Same major, same minor administrative unit	University completed	Cropped	28.10
1991	42	Female	Different major administrative unit	Less than primary completed	Urban	26.70
1991	20	Female	Different major administrative unit	Less than primary completed	Forest	24.30
1991	39	Male	Same major, same minor administrative unit	University completed	Urban	26.80
1991	77	Female	Same major, same minor administrative unit	Less than primary completed	Cropped	27.70
1991	11	Female	Same major, same minor administrative unit	Less than primary completed	Cropped	22.30
1991	31	Female	Same major, same minor administrative unit	University completed	Pasture	25.10
1991	23	Male	Same major, same minor administrative unit	Primary completed	Forest	24.40
1991	24	Female	Same major, same minor administrative unit	University completed	Urban	21.50
1991	40	Female	Same major, same minor administrative unit	University completed	Urban	23.40



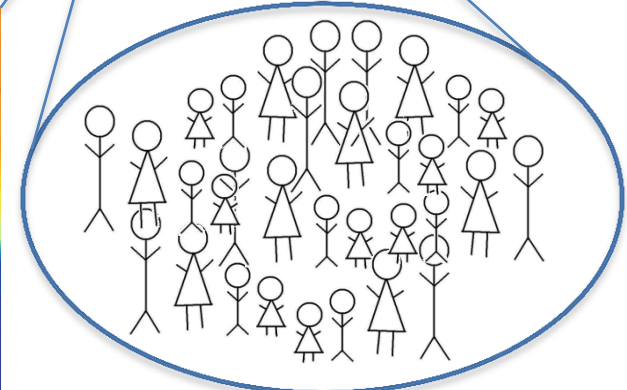
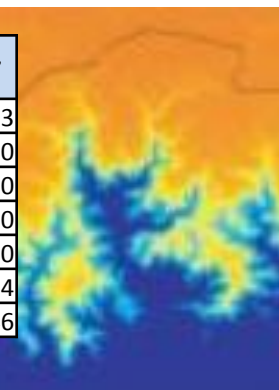
# Data Fusion – Small-Area Output

Aggregate data for administrative districts with tabulated population data and environmental characteristics

- Min/ max
- Mean
- Mode



County ID	Mean Ann. Temp.	Max. Ann. Precip.	Rent, Rural	Rent, Urban	Own, Rural	Own, Urban	Vacant, Rural	Vacant, Urban
G17003100001	21.2	768	3129	1063	637	365	34	33
G17003100002	23.4	589	2949	1075	1469	717	0	0
G17003100003	24.3	867	3418	1589	1108	617	0	0
G17003100004	21.5	943	1882	425	202	142	123	0
G17003100005	24.1	867	2416	572	426	197	189	0
G17003100006	24.4	697	2560	934	950	563	220	14
G17003100007	25.6	701	2126	653	321	215	209	46

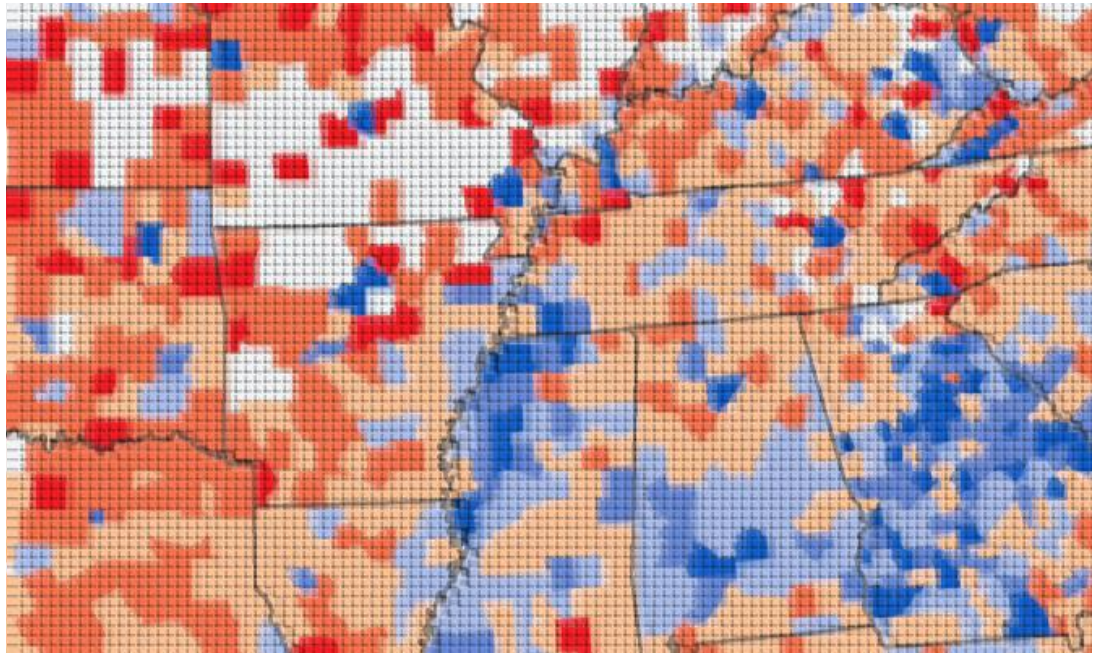




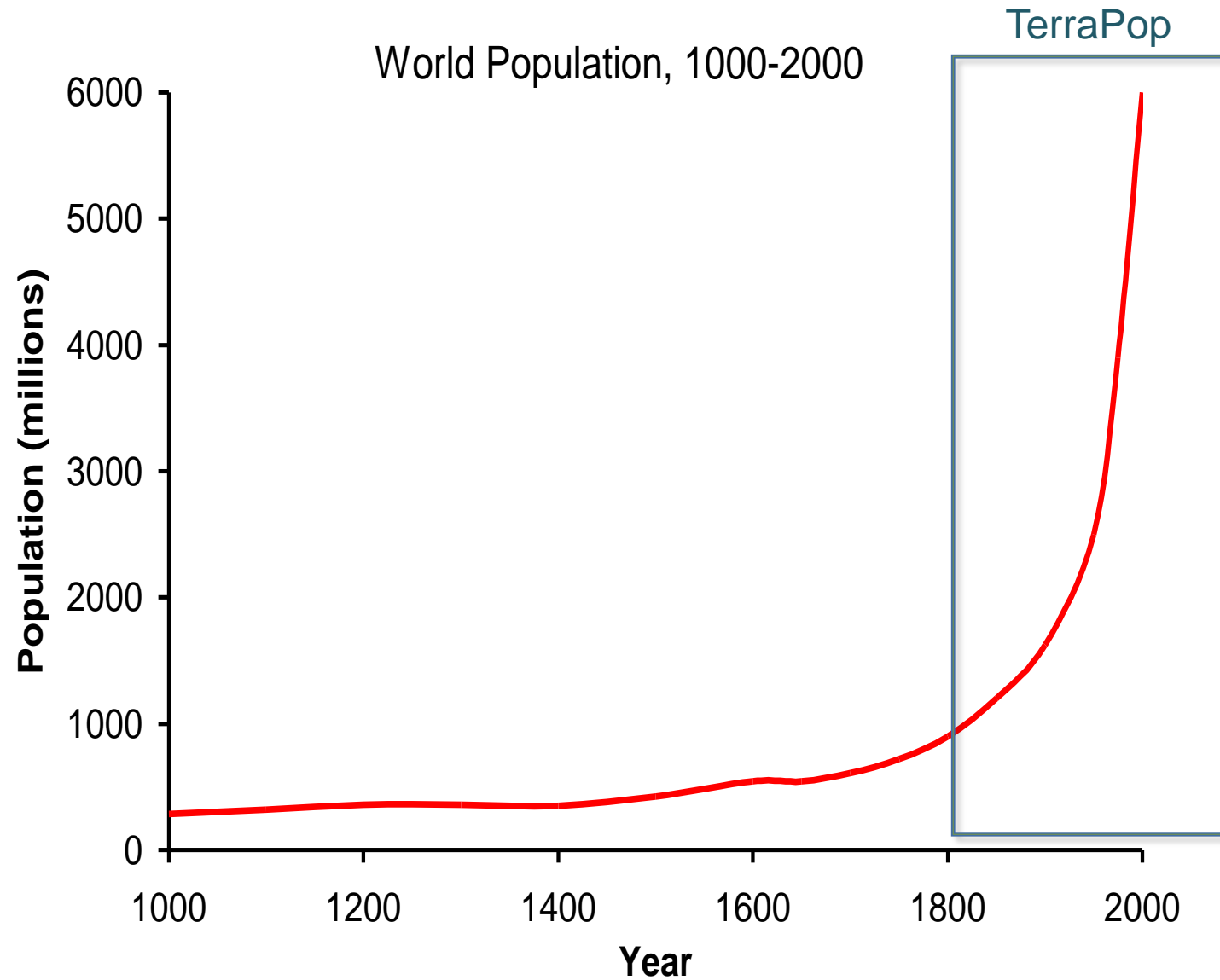
# Data Fusion – Gridded Output

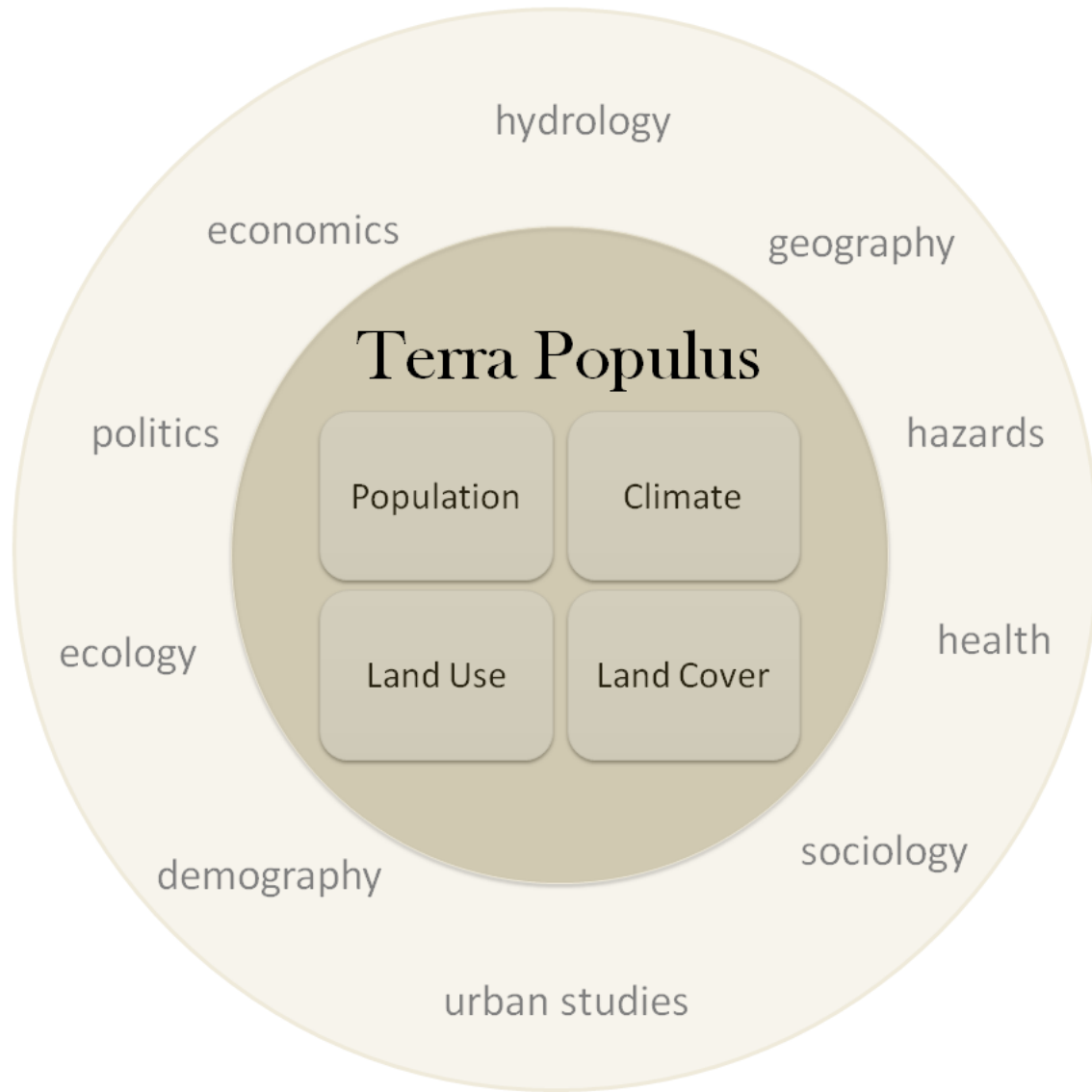
Gridded data with characteristics of population and environment

Generate  
population data in  
format compatible  
with environmental  
models



# The Temporal Dimension





# Sustainability

Create a sustainable organization that can guarantee preservation and access over multiple decades

- Organizational sustainability
- Financial sustainability
- Technological sustainability



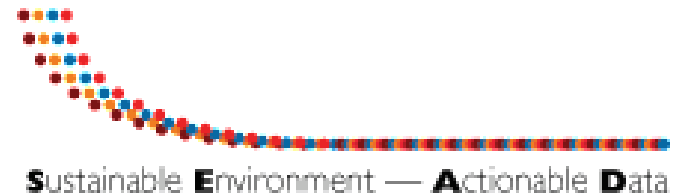




# SEAD

Sustainable Environment – Actionable Data

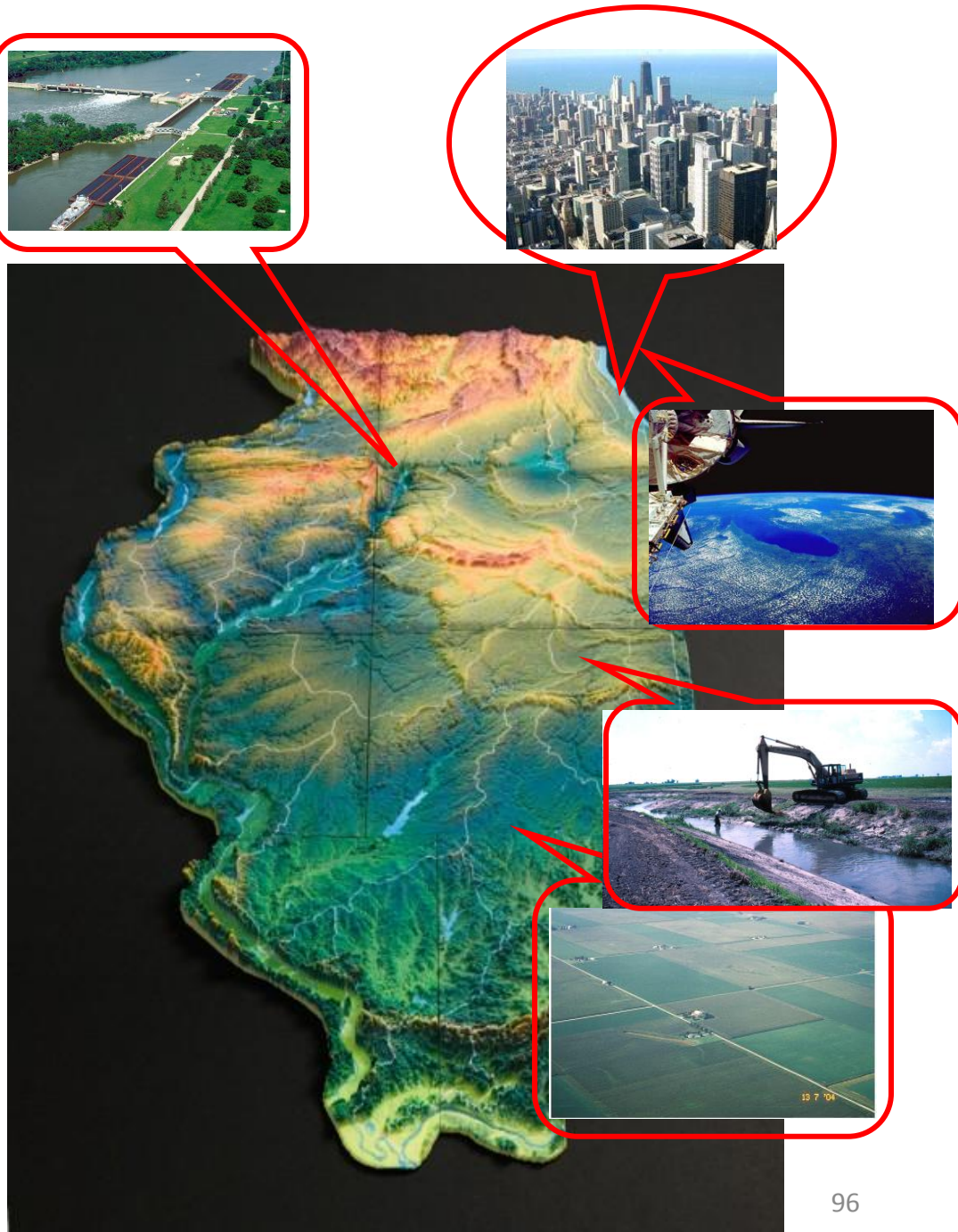
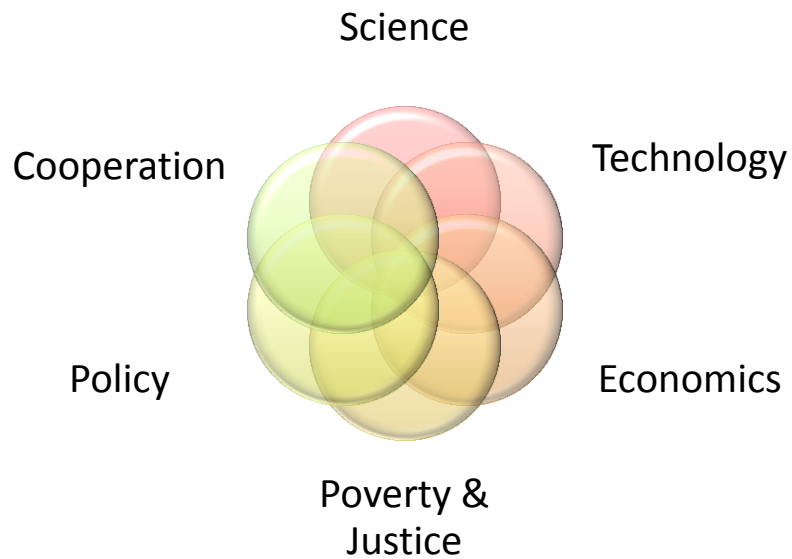
Margaret Hedstrom, PI (Michigan)  
Praveen Kumar, co-PI (Illinois)  
Jim Myers, co-PI (RPI)  
Beth Plale, co-PI (Indiana)  
Ann Zimmerman, co-PI (Michigan)



# SEAD's Goals

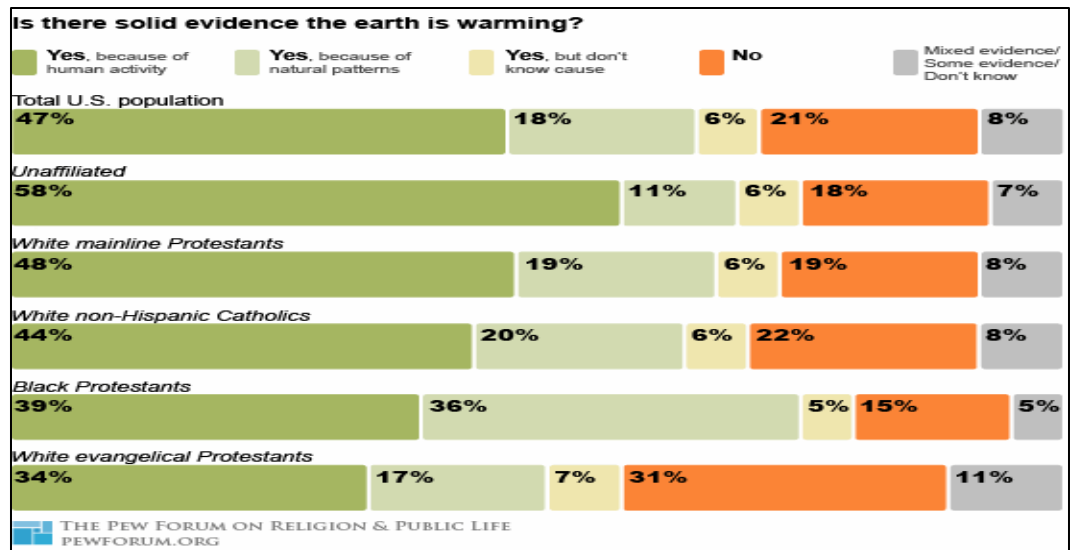
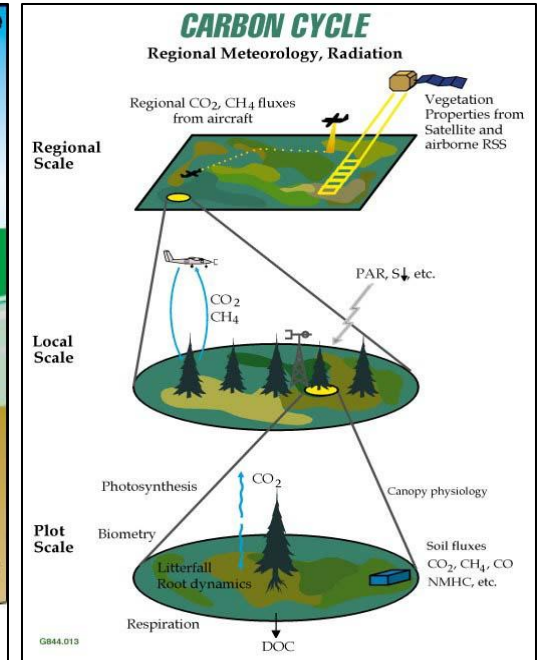
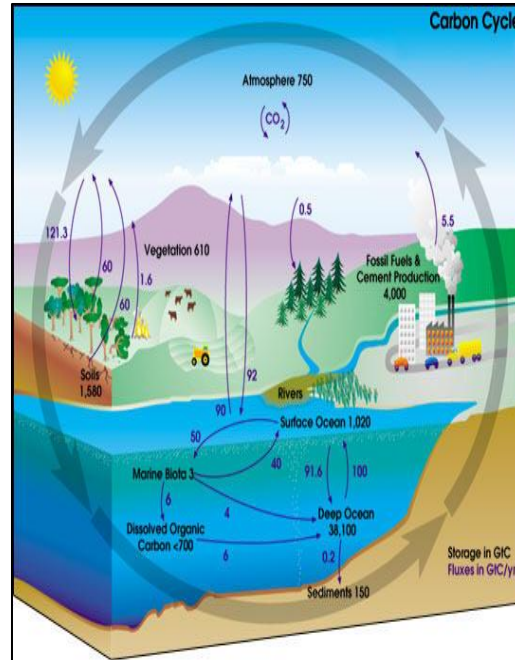
- Provide data services that address the needs of researchers in sustainability science
- Integrate these services into an generalizable “Active and Social Curation” infrastructure suited to data in the “long tail”
- Develop capabilities to package and migrate the most valuable datasets to a federated repository infrastructure for long-term preservation

# Sustainability Science



# Data challenges

- Small and derived data sets
- Heterogeneous data
- Multiple sources of data
- Short-lived data with long-term value
- Value of data grows when combined & integrated



# SEAD's Strategy

- Leverage social media for discovery of data, interest, and expertise
- Move data curation upstream in the data life cycle
- Involve domain scientists in setting priorities for evolution of data and services
- Take advantage of existing infrastructures (Institutional Repositories, ICPSR) for long-term preservation

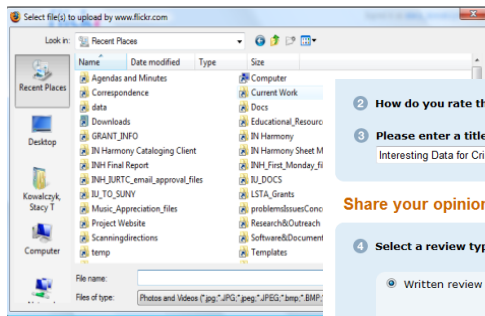


# Active Curation Model

## Active Curation

## Social Media

## Workflows



## Data

- 2 How do you rate this item? ★★☆☆
- 3 Please enter a title for your review:  
Interesting Data for Criminologists

### Share your opinion

#### 4 Select a review type

- ☒ Written review ☐ Video review

#### Type your review in the space below

[Insert a product link](#) [\(What's this?\)](#)

## Metadata

### Customer Reviews

1 Review  
2 stars (1)  
4 stars (0)  
3 stars (0)  
2 stars (0)  
1 star (0)

Average Customer Review  
★★★★★ (1 customer review)

#### Most Helpful Customer Reviews

★★★★★ Think of It as a 2-1/2 LB Stimulus Package, February 25, 2009

By Yvonne (Bethesda, MD) - [See all my reviews](#)

Ever encounter a concept so revolutionary that you wonder why every & Criminal Detection (CRC Press 2009), by Christopher Westphal, is like

Admittedly, only the few, the proud, and the discerning might dare to l hardcore law-enforcement types and techies who live and die by visua fraud prevention. Yet "Data Mining" is not only eminently readable, it c Rosetta Stone, on how to extract sense and sensibility from all the pet excellence" (one insider's waggish term for stovepiped data) around us

Seriously, if the content in Data Mining's pages (think of it as a 2-1/2 l Oprah, and a certain researcher at the Library of Congress, we could li peaceful prosperity - raising our quality of life while saving (as opposed you like anything at all about Numbers, or even NCIS (the most-viewed from Baywatch), CSI, or shows of that ilk - you owe yourself a crack a workout, and impress the heck out of onlookers. Not to mention which Memorial Fund, which speaks well for the author's motives on more tha

☐ Comment | [Pamalak](#) | Was this review helpful to you? ☐ Yes ☐ No

Share your thoughts with other customers: [Create your own review](#)

## Review Rating Commenting

#### Comments on this set

[henry](#) says:  
good idea here  
Posted 12 months ago | [permalink](#)

[Jenesis](#) says:  
You should get a photo of the cylindrical "over easy" egg that McDonald's uses for the Egg McMuffin.  
Posted 12 months ago | [permalink](#)

[Sebastian Waters](#) says:  
hahaha, very nice idea and interesting yummy photos  
Posted 12 months ago | [permalink](#)

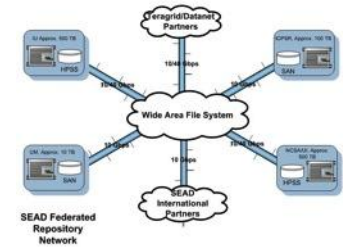
[linchongli](#) says:  
this is a great series of photos.  
Posted 12 months ago | [permalink](#)

[crystal.e](#) says:  
love this series. the cranberry sauce is so iconic.  
Posted 12 months ago | [permalink](#)

[alohu nico](#) says:  
love this. the ramen looks as though you opened the package to quickly and upended everything.

# SEAD: Leveraging Existing Resources

- Cyberinfrastructure
  - IU Data Capacitor/HPC Capabilities
  - UIUC/NCSA HPC Capabilities
  - Rensselaer CCNI Capabilities
- Repositories
  - UM Deep Blue
  - IU ScholarWorks
  - ICPSR Repository
  - UIUC IDEALS



# SEAD 18 Month Prototype Targets for Cyberinfrastructure

- Domain Engagement
  - Requirements derived from researchers
  - Use Cases
- Active and Social Content Curation
  - Pilot Active Content Repository, VIVO deployments
  - Exemplar services for Data Ingest, Discovery, Re-use, Curation
- CI for Long-term Access
  - Data model, protocol design/development
  - Pilot Federated Repository infrastructure

# SEAD TEAM

**University of Michigan:** Margaret Hedstrom (UM PI), Ann Zimmerman (Co-PI and Project Manager), George Alter, Bryan Beecher, Charles Severance, Karen Woollams, Jude Yew.

**Indiana University:** Beth Plale (IU PI), Katy Borner, Robert H. McDonald, Kavitha Chandrasekar, Robert Ping, Stacy Kowalczyk, Robert Light.

**University of Illinois:** Praveen Kumar (UIUC PI), Rob Kooper, Luigi Marini, Terry McLaren, Zaman Aktaruzzaman.

**Rensselaer Polytechnic Institute:** Jim Myers (RPI PI), Ram Prasanna Govind Krishnan, Lindsay Todd, Adam Wilson.

# Acknowledgments

SEAD is funded by the National Science  
Foundation under cooperative agreement  
#OCI0940824





# DataNet Federation Consortium

## Data Driven Science

- Implement national data grid
  - Federate existing discipline-specific data management systems to enable national research collaborations
- Enable collaborative research on shared data collections
  - Manage collection life cycle as the user community broadens
- Integrate “live” research data into education initiatives
  - Enable student research participation through control policies

Project

Shared Collection

Processing Pipeline

Digital Library

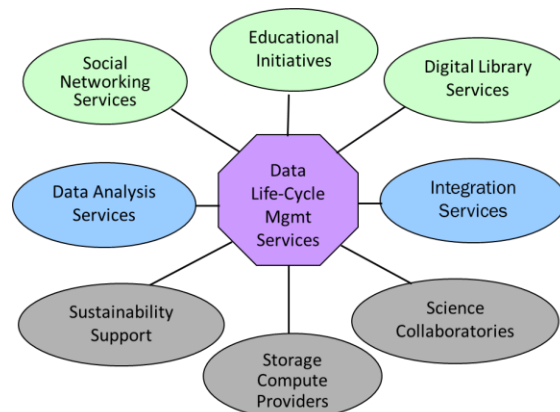
Reference Collection

Federation

*Collection Life Cycle*

### Cyber-infrastructure Partners:

Univ. of North Carolina, Chapel Hill  
Univ. of California, San Diego  
Arizona State University  
Drexel University  
Duke University  
University of Arizona  
University of South Carolina



### Science and Engineering Initiatives:

Ocean Observatories Initiative  
the iPlant Collaborative  
CUAHSI  
CIBER-U  
Odum Social Science Institute  
Temporal Dynamics of Learning Center

# Collaboration Environment

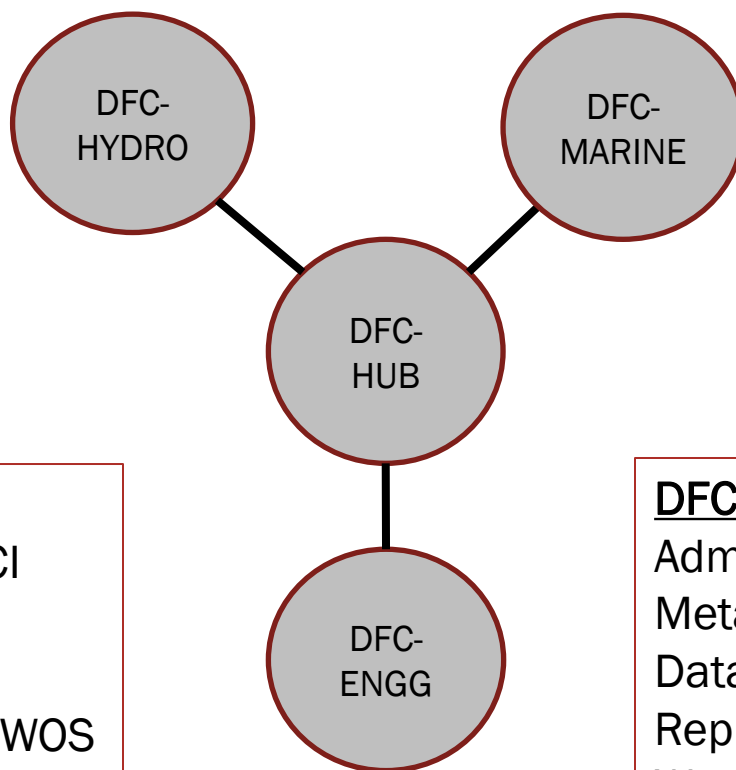
- Manage data and workflows for reproducible research
  - Enforce domain policies
  - Capture workflow process provenance
  - Interact with remote data resources
  - Capture copies of all research data and metadata
  - Orchestrate internal and external services
- Enable project to manage results
  - Create / share / publish / archive / repurpose
  - Reference collections

# DFC Grid Phase-1

## (current federation)

### DFC-Hydrology

Administration: RENCi  
 Metadata: RENCi  
 Data Resc: USC  
 Data Resc: NCDC  
 Replica Resc: RENCi  
 Workflow Resc: ALL  
 LifeCycle Engine: RENCi  
 Message Hub: RENCi



### DFC-Marine

Administration: UCSD  
 Metadata: UCSD  
 Data Resc: UCSD  
 Replica Resc: RENCi  
 Workflow Resc: ALL  
 LifeCycle Engine: UCSD  
 Message Hub: UCSD

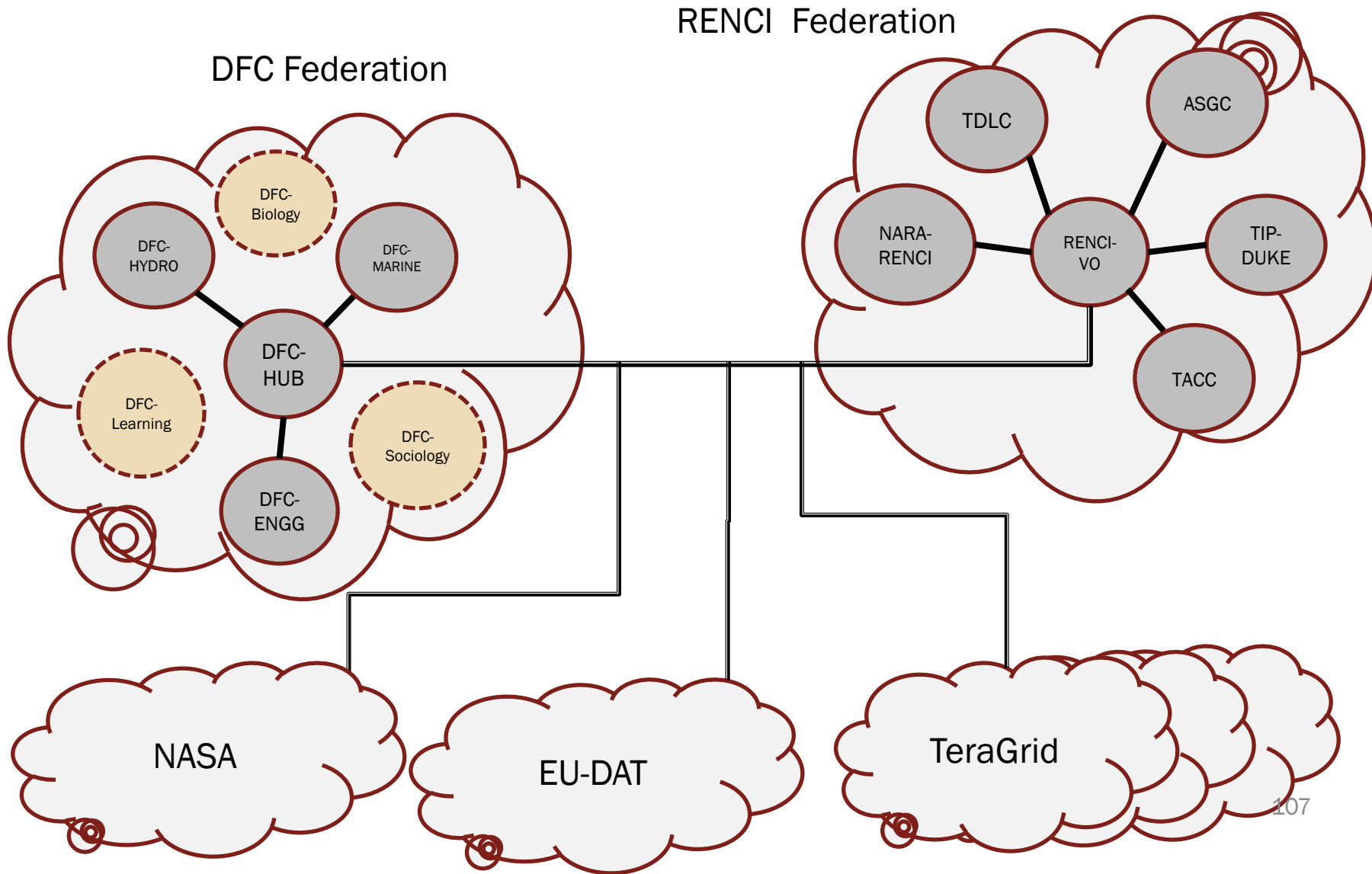
### DFC-Federation Hub

Administration: RENCi  
 Metadata: RENCi  
 Data Resc: RENCi  
 Replica Resc: RENCi-WOS  
 Workflow Resc: ALL  
 LifeCycle Engine: RENCi  
 Message Hub: RENCi

### DFC-Engineering

Administration: Drexel  
 Metadata: Drexel  
 Data Resc: Drexel  
 Replica Resc: RENCi  
 Workflow Resc: ALL  
 LifeCycle Engine: RENCi  
 Message Hub: Drexel

# Federation of Federations

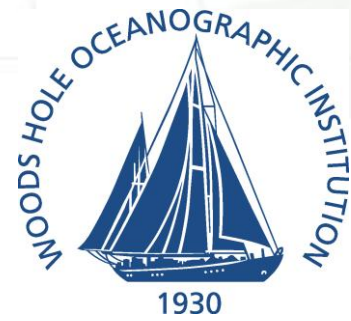




# Facilitating Next Generation Science Collaboration: Respecting and Mediating Vocabularies with Semantics in Ecosystems Assessments.

*January 26, 2012 Data 2012*

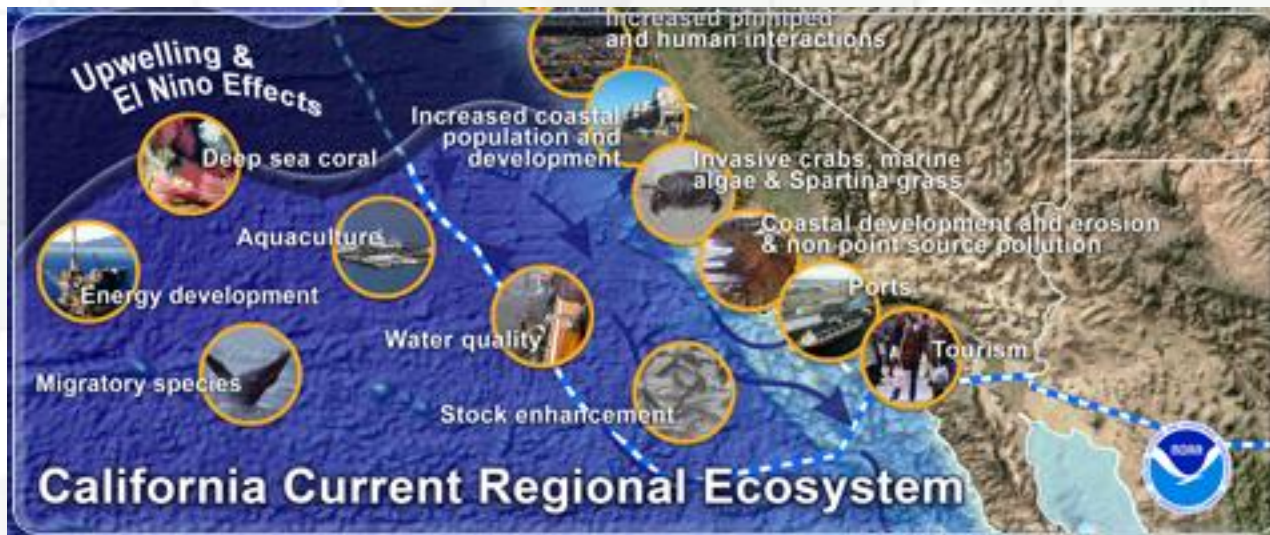
Peter Fox (RPI/ Tetherless World Constellation and WHOI/AOP&E) [pfox@cs.rpi.edu](mailto:pfox@cs.rpi.edu) and Andrew Maffei (WHOI/C&IS) [amaffei@whoi.edu](mailto:amaffei@whoi.edu)  
NSF INTEROP ECO-OP project. <http://tw.rpi.edu/web/project/ECOOP>





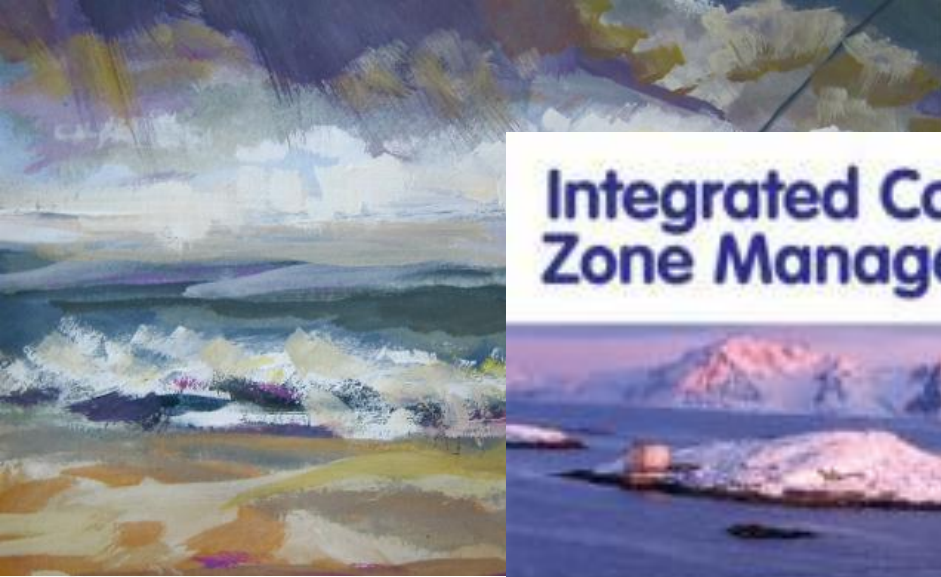


# Marine ecosystems





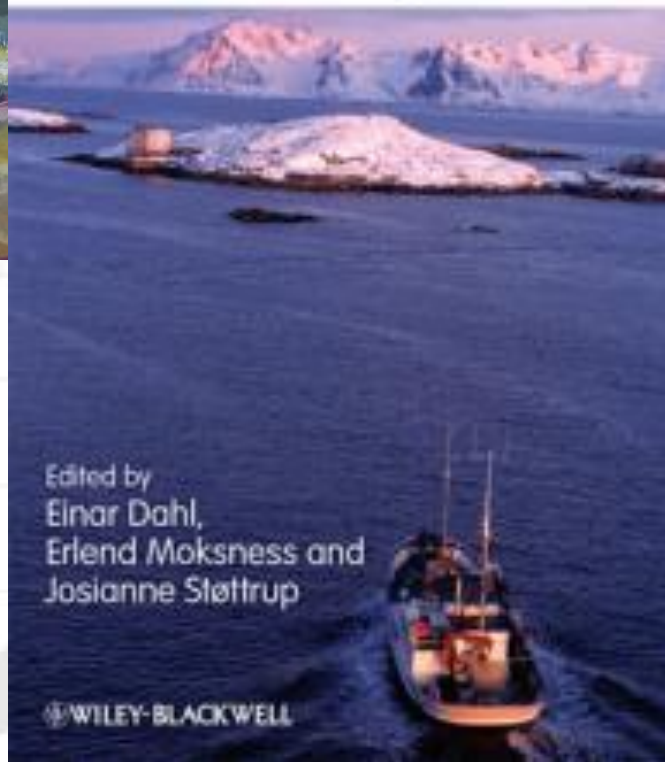
# Fish, science, decision



## Integrated Coastal Zone Management

Edited by  
Einar Dahl,  
Erlend Moksness and  
Josianne Støttrup

WILEY-BLACKWELL





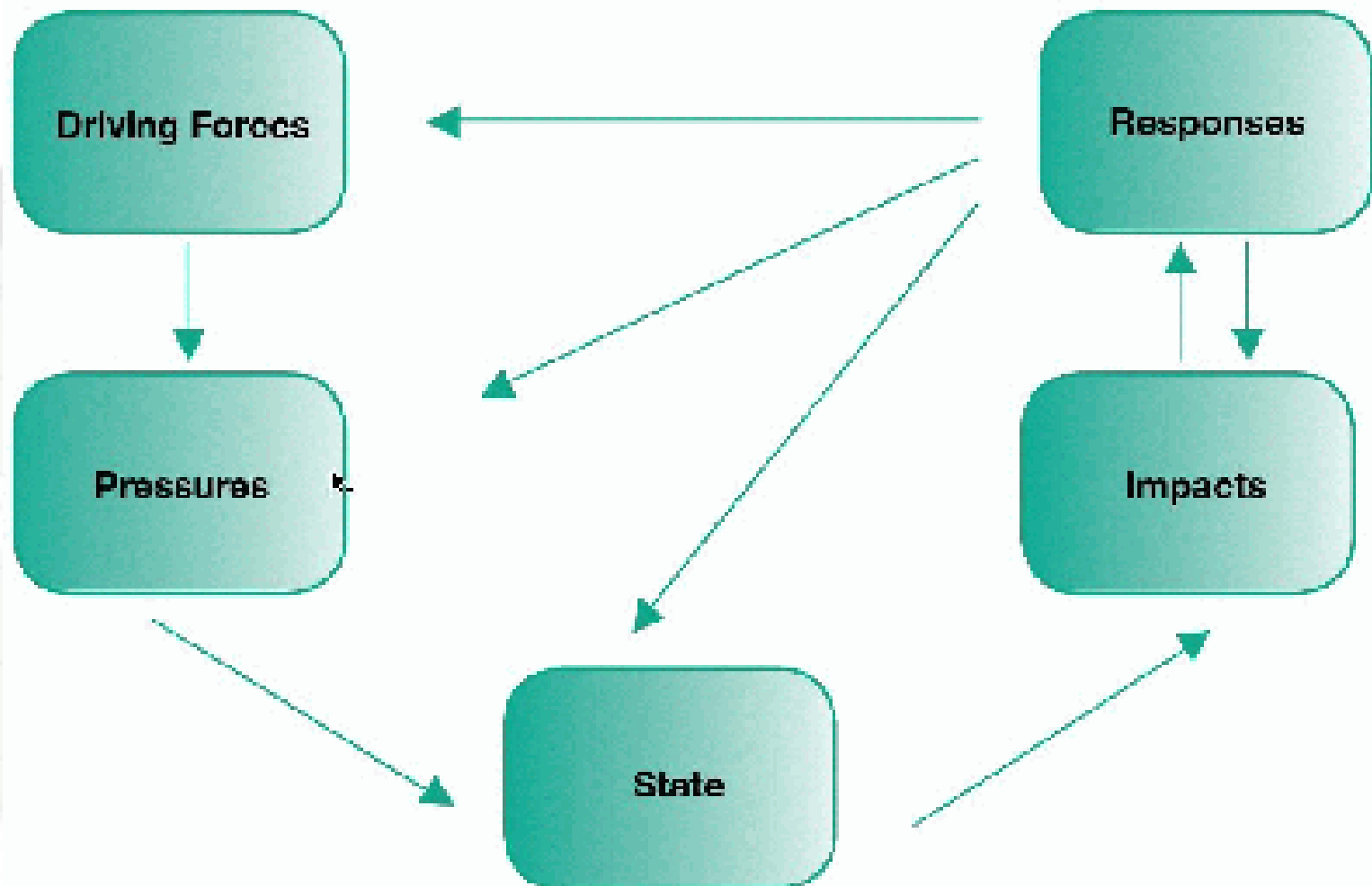


# Vision?

- “Our vision is to develop, facilitate, and maintain sustained multi-way engagement of natural and social scientists and many practitioners in multi-scale local to global networks for LMEs”.
  - Organization is required so participants can carry out their (respective) mission(s)
  - Those participants (by defn.) will never be in a single organization -> virtual organization
- Goal: We want to perform routine assessments of LMEs involving all (or as many) stakeholders and we want robust science data presented in forms that various end-users can consume...



# Framework - DPSIR





# Semantics of DPSIR?

- ▶ ● Driving force of energy development
- ▶ ● Pressure of energy development
- ▶ ● State change of energy development
- ▶ ● Impact of energy development
- Response for sustainable energy development
- Society
- Environment
- Economy
- Energy system
- Energy supply sector
- Energy service
- ▶ ● Energy technology
- ▶ ● Energy resource
- ▶ ● Energy supply process

- Pressure of energy development
- ▶ ● GHG emission from energy production and use
- Contaminant discharges in liquid effluents
- ▶ ● Air pollutants emission from energy systems
- Generation of radioactive wastes
- Accident as a result of energy uses
- Generation of wastes
- Land area taken up by energy facilities
- Radionuclides in atmospheric radioactive discharges
- Oil discharges into coastal waters
- Forests' income to energy prices
- ▶ ● Energy resources depletion





# Drivers/ Pressures

- ***Physical Drivers***

- North Atlantic Oscillation
- Atlantic Multi-decadal Oscillation

- ***Human Drivers***

- Population
- Income

- ***Human Pressures (Fishery Removals)***

- Number Groundfish Vessels
- Landings, Principal Groundfish
- Landings, Other Fish
- Landings, Small Pelagics
- Landings, Crustaceans
- Landings, Molluscs

- ***Temperature***

- Extended Reconstructed SST
- Coastal Temperature, Virginia
- Coastal Temperature, Woods Hole
- Coastal Temperature, Boothbay Harbor
- Survey sea surface temperature
- Survey bottom sea temperature
- Thermal Habitat <4°C
- Thermal Habitat >5°C and <15 °C
- Thermal Habitat >16°C

- ***River Discharge***

- River Flow-Gulf of Maine
- River Flow-Middle Atlantic Bight
- River Flow-Southern New England

- ***Wind Fields***

- Wind Stress, Cape Hatteras
- Wind Stress, New York
- Wind Stress, Georges Bank
- Wind Stress East-West, Cape Hatteras
- Wind Stress East-West, New York
- Wind Stress East-West, Georges Bank
- Wind Stress North-South, Cape Hatteras
- Wind Stress North-South, New York
- Wind Stress North-South, Georges Bank

- ***Other***

- Stratification
- Survey surface salinity
- Survey bottom salinity
- Gulf Stream Location
- %Labrador-Subarctic Slope Water in GoM



# *Ecosystem State Variables*

## *Plankton*

- Continuous Plankton Recorder Color Index.
- Zooplankton Ecosystem Biovolume
- Ratio of Small to Large Zooplankton

## *Nekton/Benthos*

- Relative Abundance, Crustaceans
- Relative Abundance, Elasmobranch
- Relative Abundance, Ground Fish
- Relative Abundance, Molluscs
- Relative Abundance, Other Fish
- Relative Abundance, Small Pelagics
- Relative Abundance, All Species

## *Demography/Trophic Level*

- Mean Trophic Level Catch
- Mean Trophic Level Survey
- Primary Production Required, Landings
- Mean Length

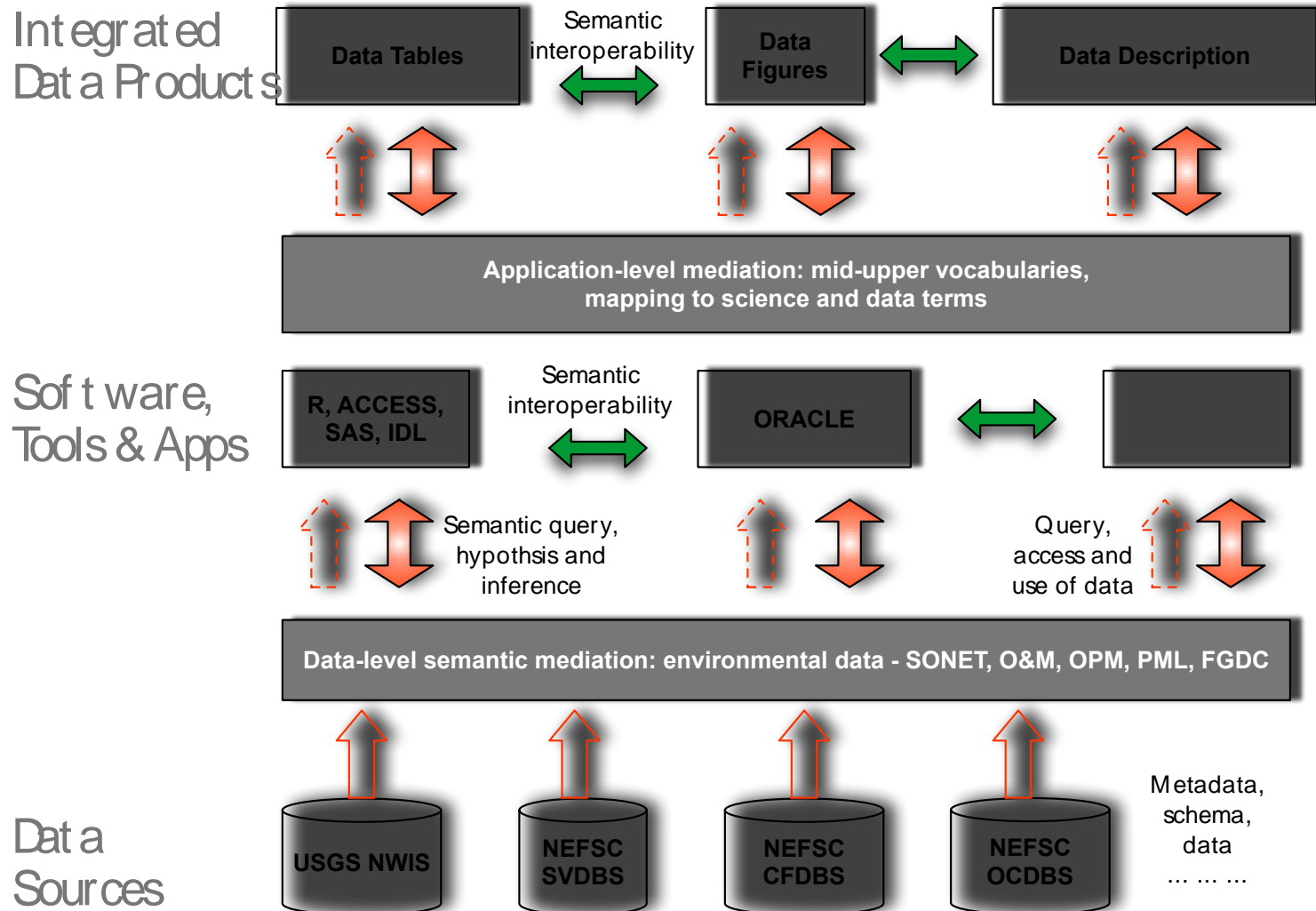
## *Community Composition*

- Thermal Preference
- Pelagic to Demersal Ratio
- Elasmobranch to Demersal Groundfish Ratio
- Impacts
- Groundfish Fishery Revenue



# Northeast Status Report

## Ecosystem Status Report



# **Ecosystem Status Report**

**For the Northeast U.S. Continental Shelf Large Marine Ecosystem**



Ecosystem Assessment Program, Northeast Fisheries Science Center

July 2009

## **Main Findings**

The Northeast U.S. Continental Shelf Large Marine Ecosystem (NES LME) has undergone sustained perturbations due to environmental and anthropogenic impacts over the last four decades, resulting in fundamental changes in system structure.

Thermal conditions in the NES LME are changing due to warming of coastal and shelf waters and cooling in the northern end of the range. As a consequence, there has been a constriction of thermal habitats in the ecosystem, a northward shift in the distributions of some fish species and a shift to a warmer-water fish community.

Zooplankton community structure has also changed in concert with climate and physical processes acting over the North Atlantic Basin indicating the importance of remote forcing to the function and structure of this ecosystem

Important changes in some components of benthic communities, notably increased abundance of sea scallops and lobster are evident, reflecting changes in fishery management and/or ecological conditions.

The direct and indirect effects of species-selective harvesting patterns have also contributed to shifts fish community composition which is now dominated by small pelagic fishes and elasmobranch species (skates and small sharks) of low relative economic value.

The trajectory of regional human population size suggests that anthropogenic pressure in the ecosystem will continue to increase.

The Northeast U.S. Continental Shelf is classified as experiencing ecosystem overfishing according to published criteria for this designation, although improvement in the condition of several resource species has occurred and exploitation effects have been reduced for some system components over the last decade.



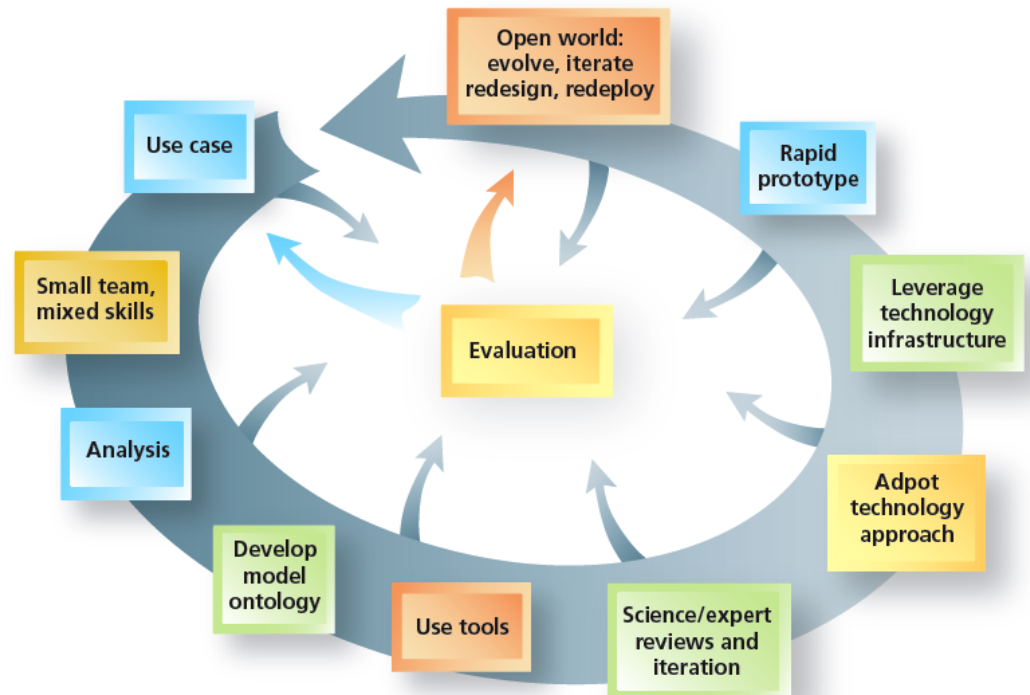


# Informatics enables a new approach

- Use cases
- Stakeholders
- Distributed authority
- Access control
- Ontologies
- Maintaining Identity

## Semantic Web Methodology & Technology Development Process

- ▶ Establish and improve a well-defined methodology vision for semantic technology based on application development
- ▶ Leverage controlled vocabularies, etc.







## SONet:

A Community-Driven **Scientific Observations Network** to  
achieve Semantic Interoperability of Environmental and  
Ecological Data  
(NSF OCI-INTEROP 0753144)

M. Schildhauer<sup>1</sup>, L. Bermudez<sup>2</sup>, S. Bowers<sup>3</sup>, M. Jones<sup>1</sup>,  
Steve Kelling<sup>4</sup>, Hilmar Lapp<sup>5</sup>, **Deborah McGuinness**<sup>6</sup>

1. NCEAS, Santa Barbara, CA

2. Open Geospatial Consortium, MA

3. Gonzaga University, WA

4. Cornell University, Ithaca, NY

5. NESCent, Durham, NC

6. Rensselaer Polytechnic Institute, Albany, NY



Data 2012, Indianapolis

# Observational data

*Much earth and life science data consists of*  
*OBSERVATIONS:*

- *Measurements* (categorical, numerical) of some
- *Characteristics* (attributes, properties) of some
- *Entity* (specimen, phenomenon, “thing”)

# Motivation

Many “semantic” efforts in earth/biodiversity/environmental sciences, exploring use of ***observational construct as foundational template for organizing data***

Specialized concerns of different domains may drive semantic solutions to be ***diverse and incompatible***

Opportunity for communicating among different domains to achieve ***greater interoperability*** of emerging semantic technology solutions through a ***shared core model for observations***

# Observational data models...

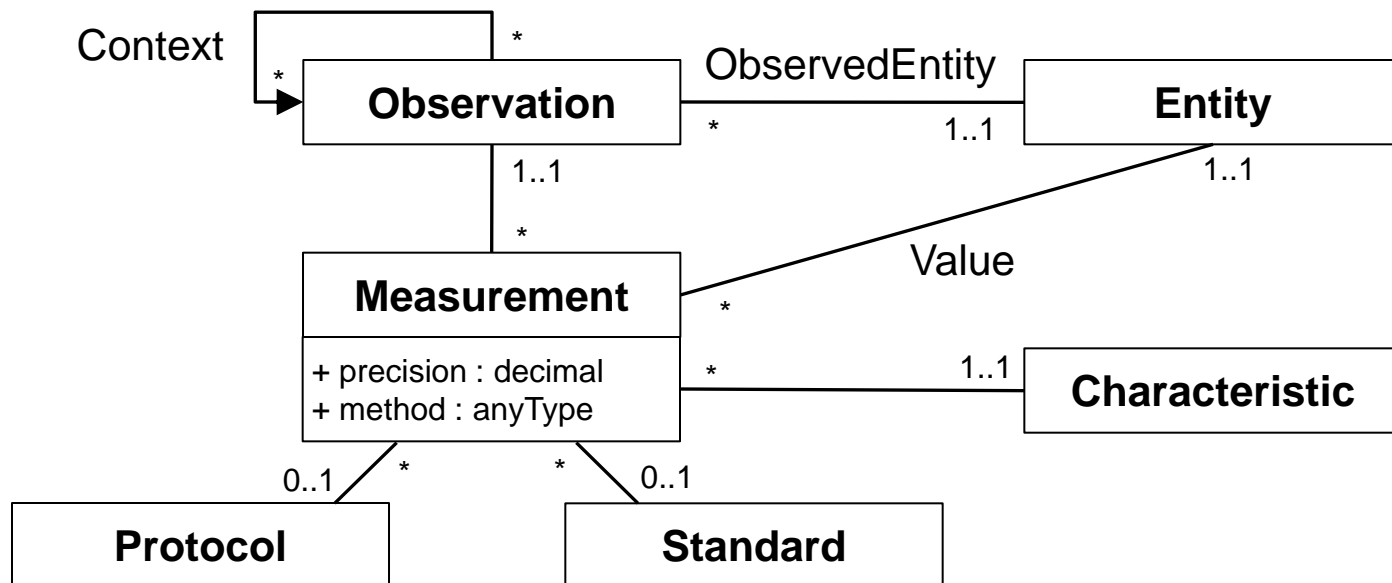
Project	Domain	Observational data model
VSTO	Atmospheric sciences	Ontologies for interoperability among different meteorological metadata standards and other atmospheric measurements
SERONTO	Socioecological research	Ontology for integrating socio-ecological data
OGC's O&M	Geospatial	Observations and Measurements standard for enhancing sensor data interoperability
SEEK's OBOE	Ecology	Extensible Observation Ontology for describing data as observations and measurements
PATO's EQ	Phenotype/Evolution	Underlying model for describing phenotypic traits to link with genomic data

# Formalizing the Observational Data Model



*SONet*

- Implemented as an OWL-DL ontology
  - Provides basic concepts for describing observations
  - Specific “extension points” for domain-specific terms

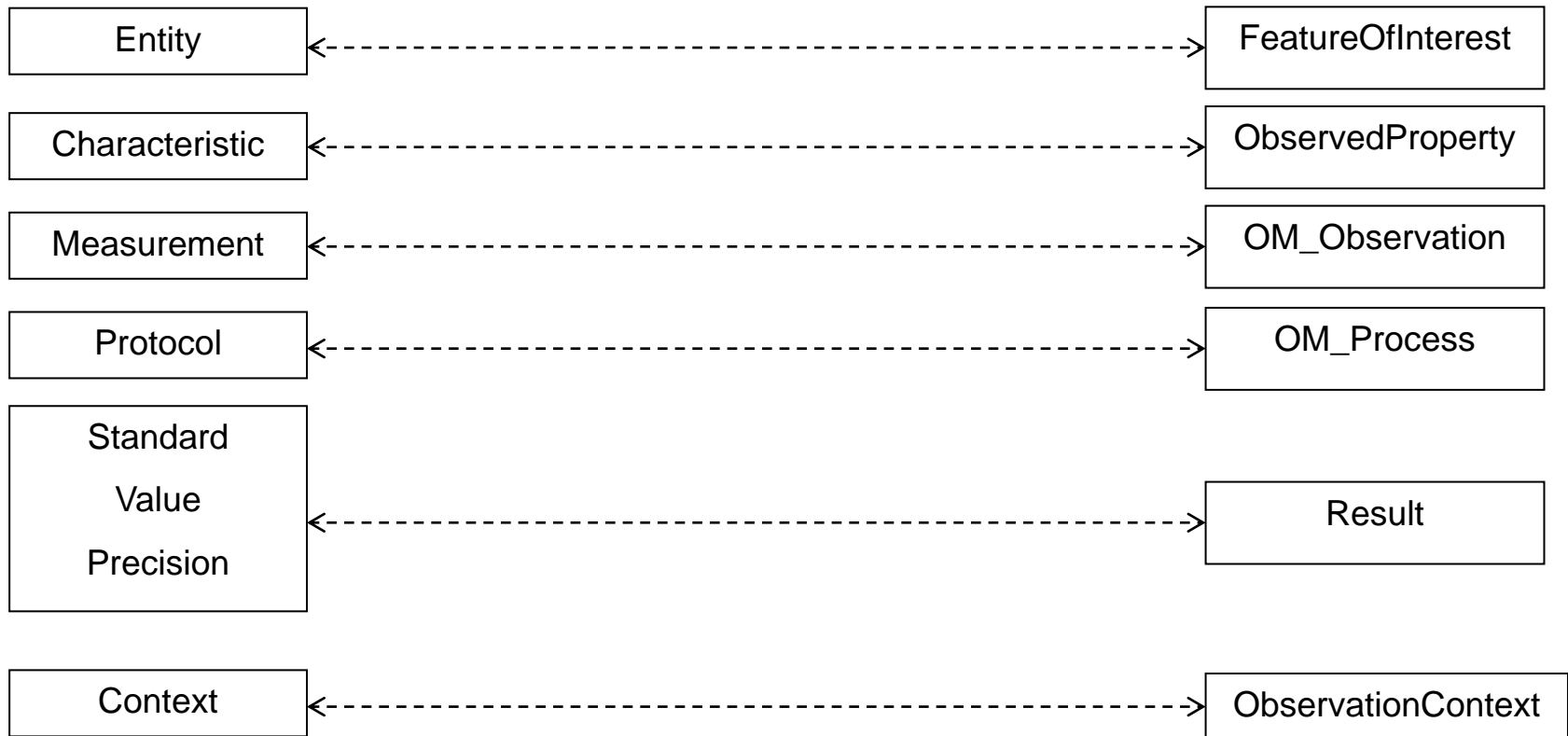




# Similarities among Observational Data Models

## SONet's OBOE Extensible Observation Ontology

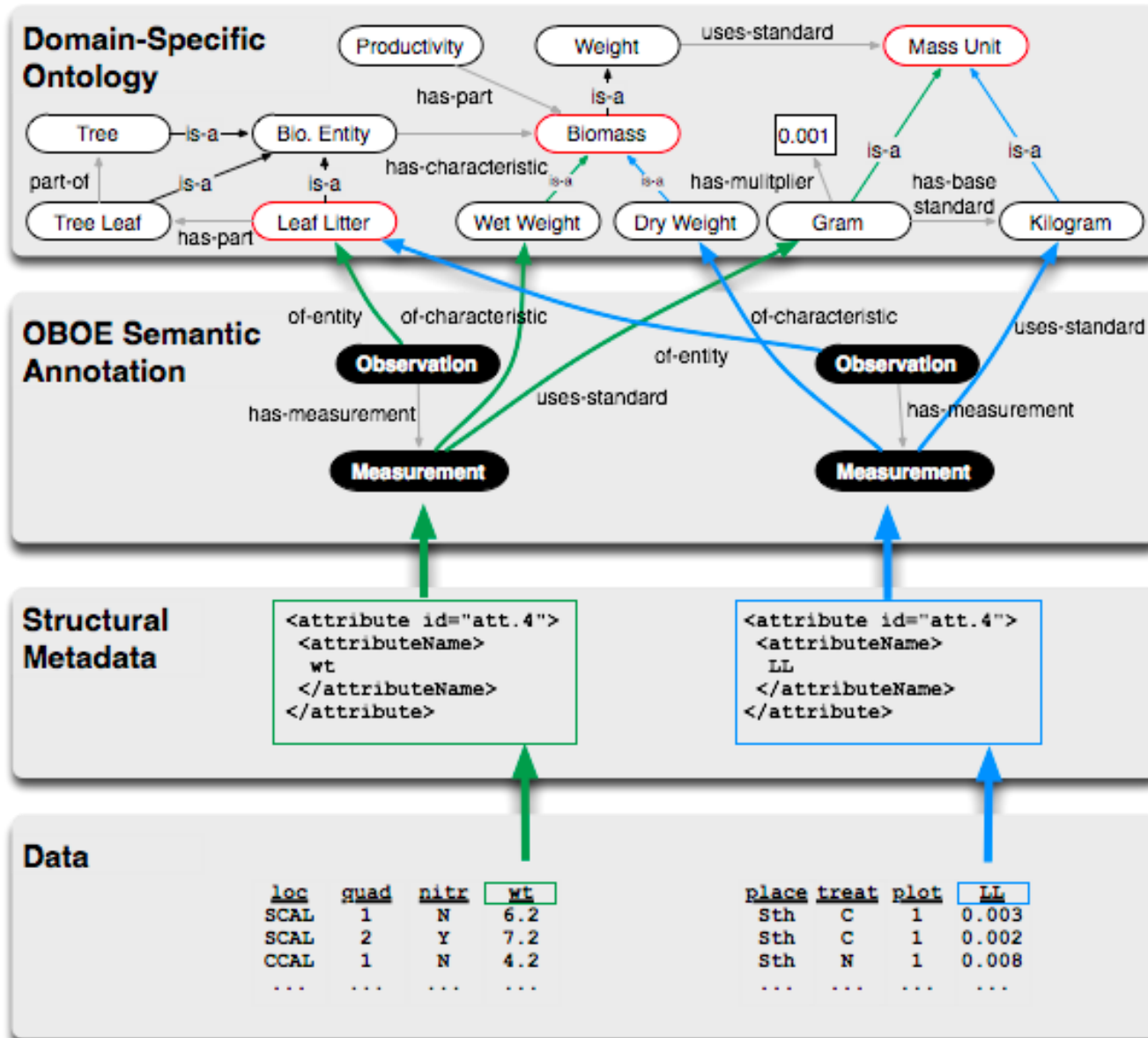
## OGC's Observations and Measurements



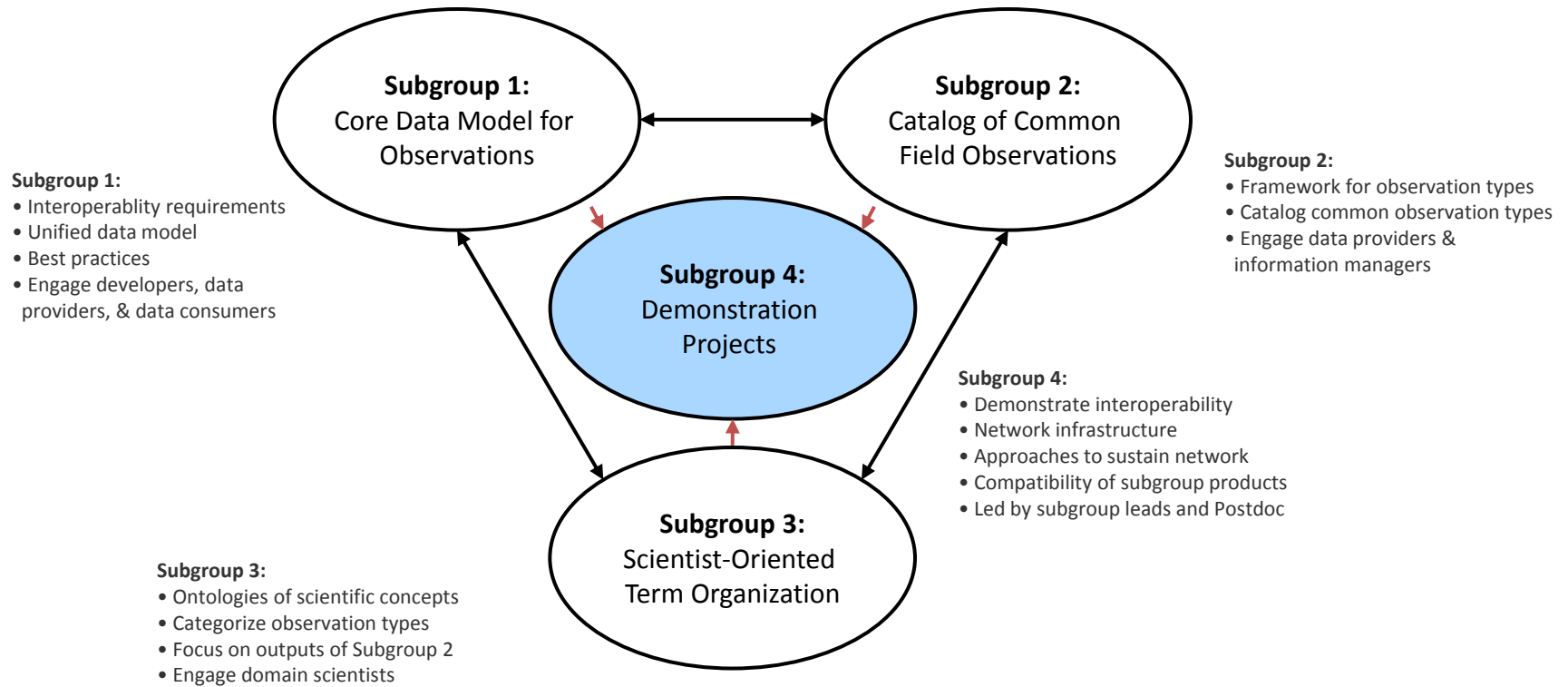
# Semantic Annotation: linking data values to concepts through observations

- Observational data models provide high-level, domain-neutral abstraction of scientific observations and measurements
- Link data (or metadata) through observational data model to *terms from domain-specific ontologies*
- *Context* can inter-relate values in a tuple
- Provide clarification of semantics of data set as a whole, not just “independent” values or single attributes

# Semantic annotation



# SONet Activities

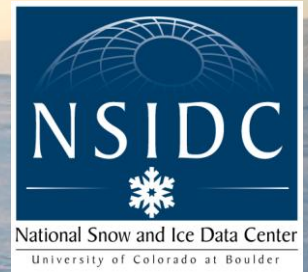


# Selected Synergies



- *Semtools effort (NSF ABI)*
  - *Extending Morpho metadata editor to permit semantic annotation; close links with KNBrepository that is an initial node on DataONE*
- *OGC Observations and Measurements*
  - *Working with OGC developers to assure compatibility with the OGC/ISO “O&M” specification (in XMLS)*
- *DataONE Semantics Working Group (NSF Datamet)*
  - *Cross-project participants from SONet involved in discussing semantics approaches for use in DataONE (McGuinness co-chair; Schildhauer, Bermudez, Lapp members)*
  - *Questions / info: [d1m@cs.rpi.edu](mailto:d1m@cs.rpi.edu) or [schild@nceas.ucsb.edu](mailto:schild@nceas.ucsb.edu)*
  - <http://sonet.ecoinformatics.org>





# The Semantic Sea Ice Interoperability Initiative

---

Mark A. Parsons, Ruth Duerr, Siri Jodha Singh Khalsa, Peter Pulsifer  
National Snow and Ice Data Center, Univ. Colorado

Peter Fox, Deborah McGuinness, James McCusker  
Tetherless World Constellation, RPI



Image courtesy Andy  
Mahoney, NSIDC



# SSI works to make Arctic data more useful to more people.

Extend a network of Arctic data and systems and harmonize metadata.  
Create integrative sea ice ontologies and encourage their use.  
Improve the discovery, understanding, and use of sea ice data.

Photo courtesy Ted  
Scambos, NSIDC



Sea Ice Index  
Sea Ice Extents, 1979-2006

monthly median sea ice extent 1979-2000

1979 Sept



Image © 2007 NASA



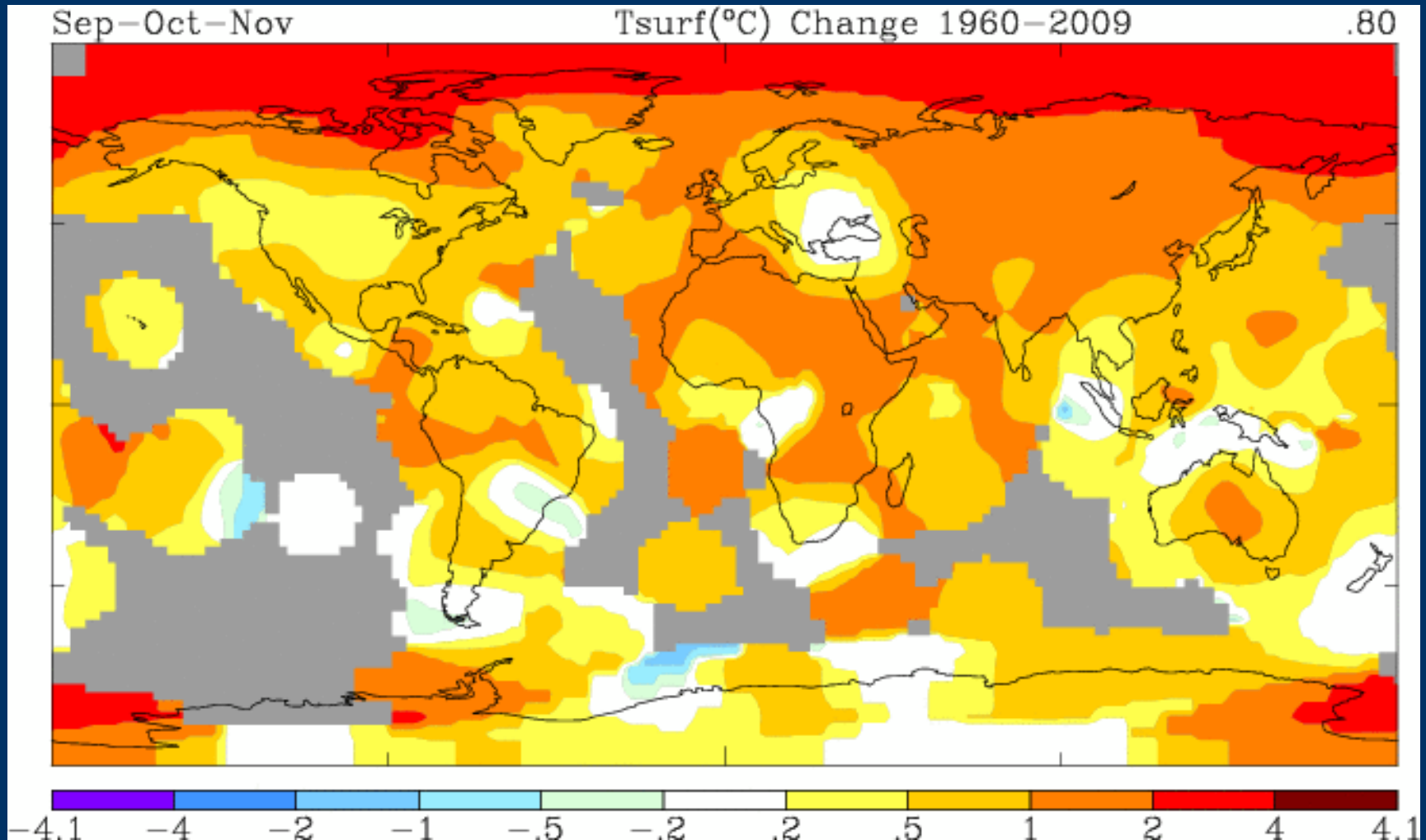
Google

September 2007



Google

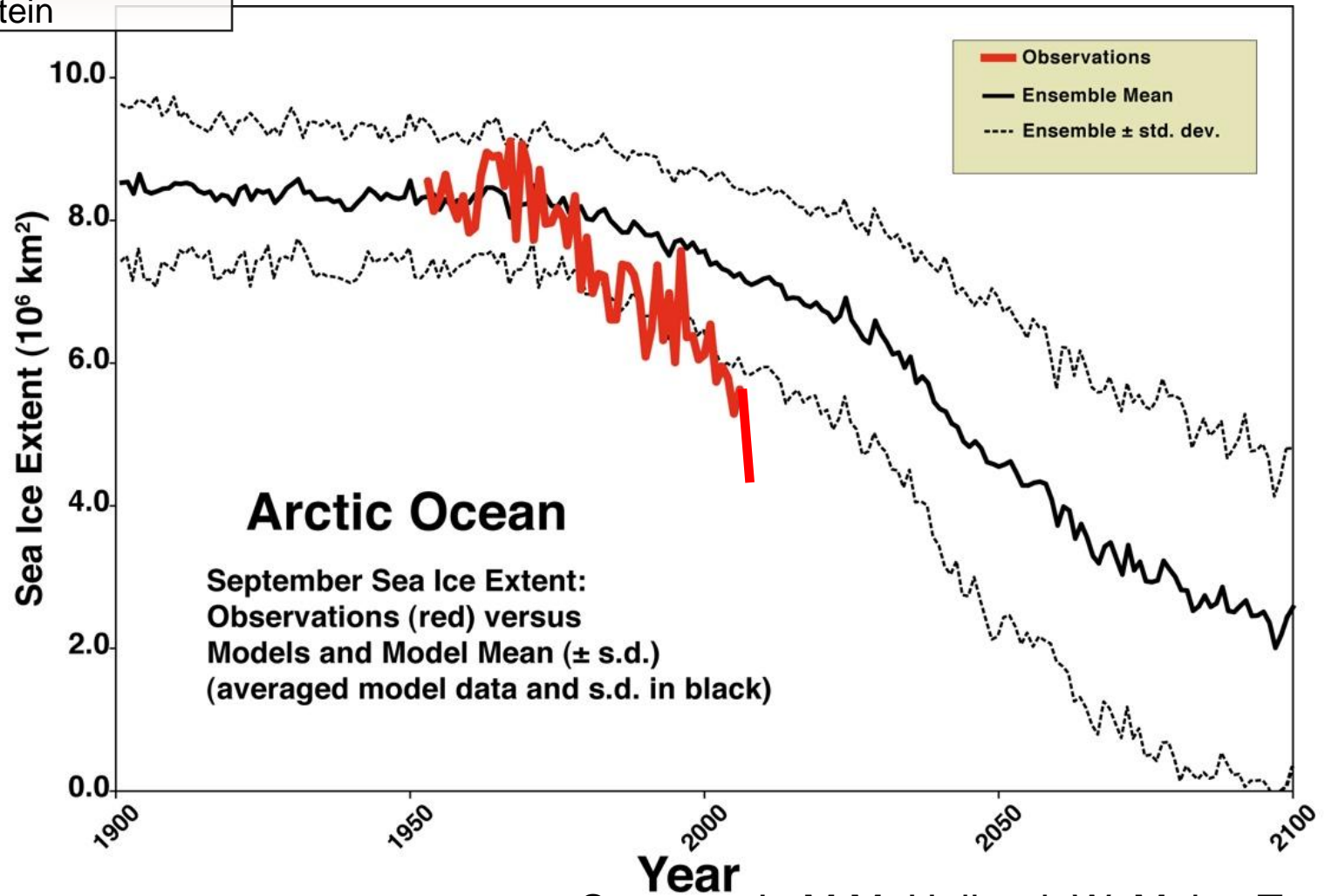
# Observed autumn temperature trends, 1960-2009



GISS Analysis and NCEP/NCAR Reanalysis (Courtesy M. Serreze)



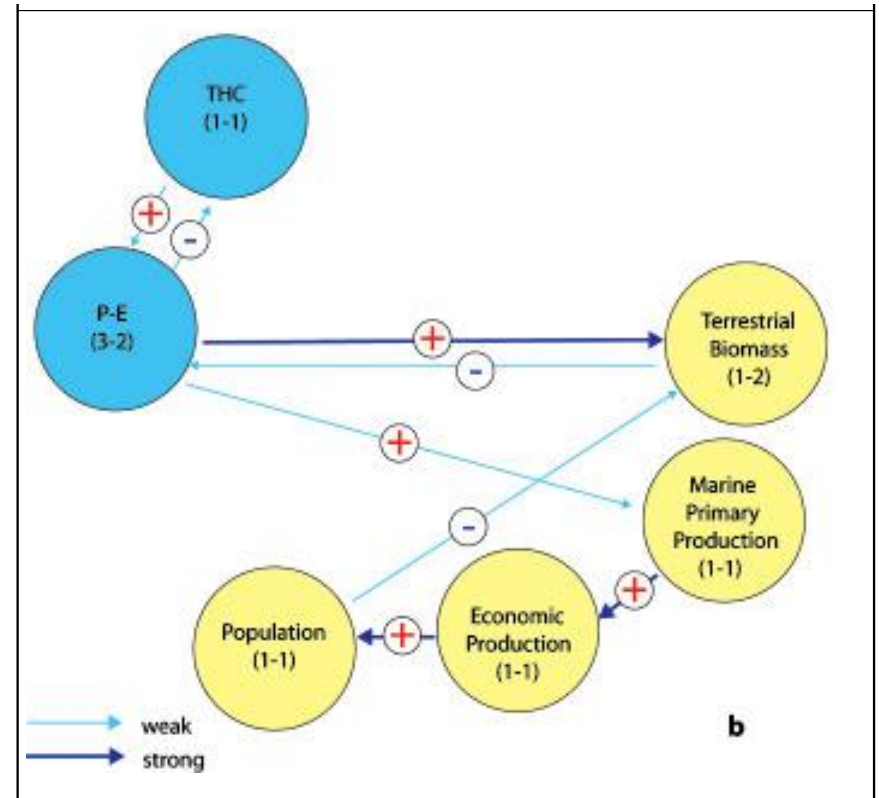
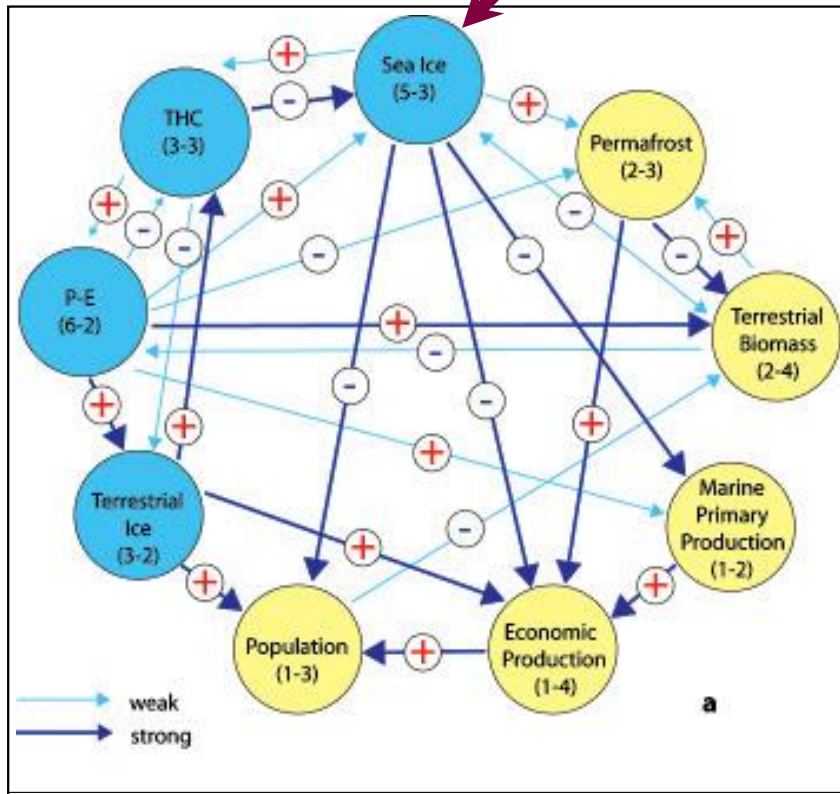
If we knew what it was we  
were doing, it would not be  
called research, would it?  
- Albert Einstein



Stroeve, J., M.M. Holland, W. Meier, T.  
Scambos, M. Serreze, 2007.



sea ice



The Arctic system with and with out sea ice from Overpeck et al. 2005.



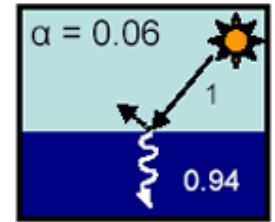
photo courtesy *The Inquisitr*

## Operational Perspective

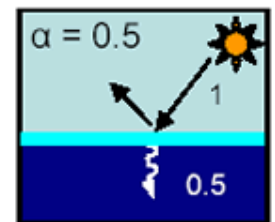
A Coast Guard Icebreaker cutting a path for a tanker to get fuel to the iced in city of Nome, Alaska.



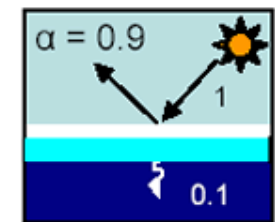
I. Open ocean



II. Bare ice



III. Ice with snow



© Karen Frey, The Polaris Project

## Research Perspective

How is the changing sea ice affecting the ice-albedo feedback?





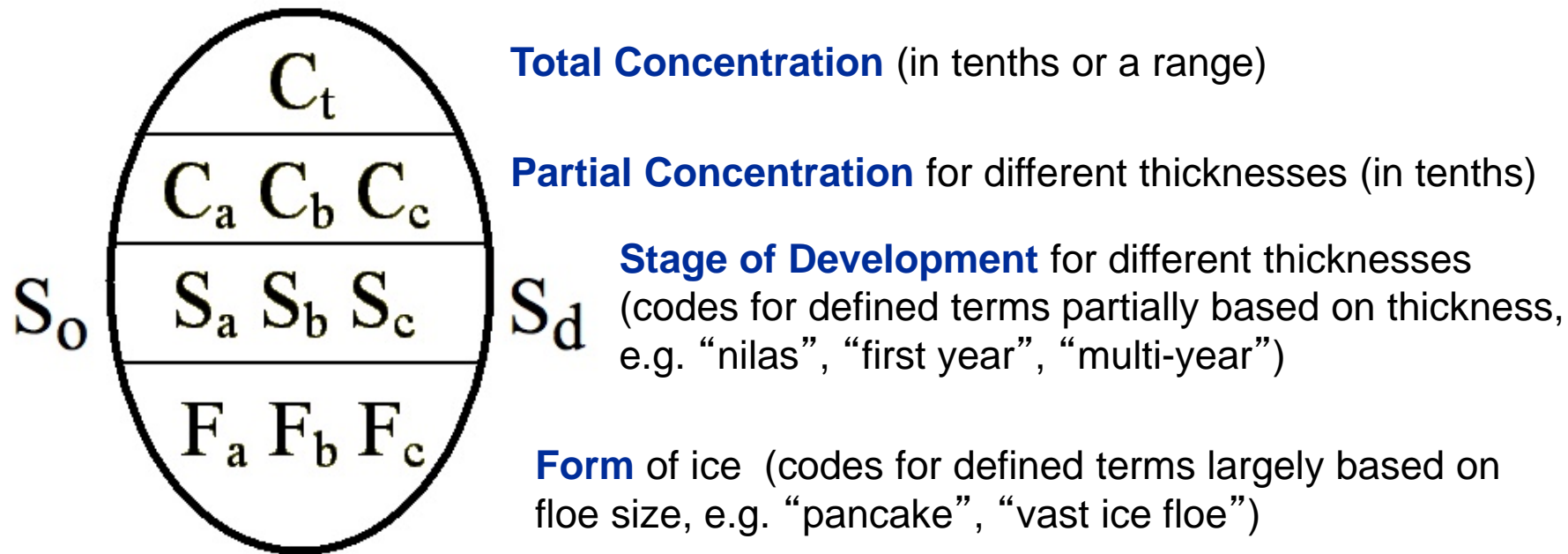
# The Local Indigenous Perspective

- earlier break up / later freeze up (2-3 weeks each)
- increased weather variability / traditional forecasts no longer work
- sea ice thinner; poorly formed (poor strength/integrity)
- seasonal calendar off; some names no longer apply
- etc. etc.

photo ©Shari  
Gearheard

# The WMO “Egg Code”

---



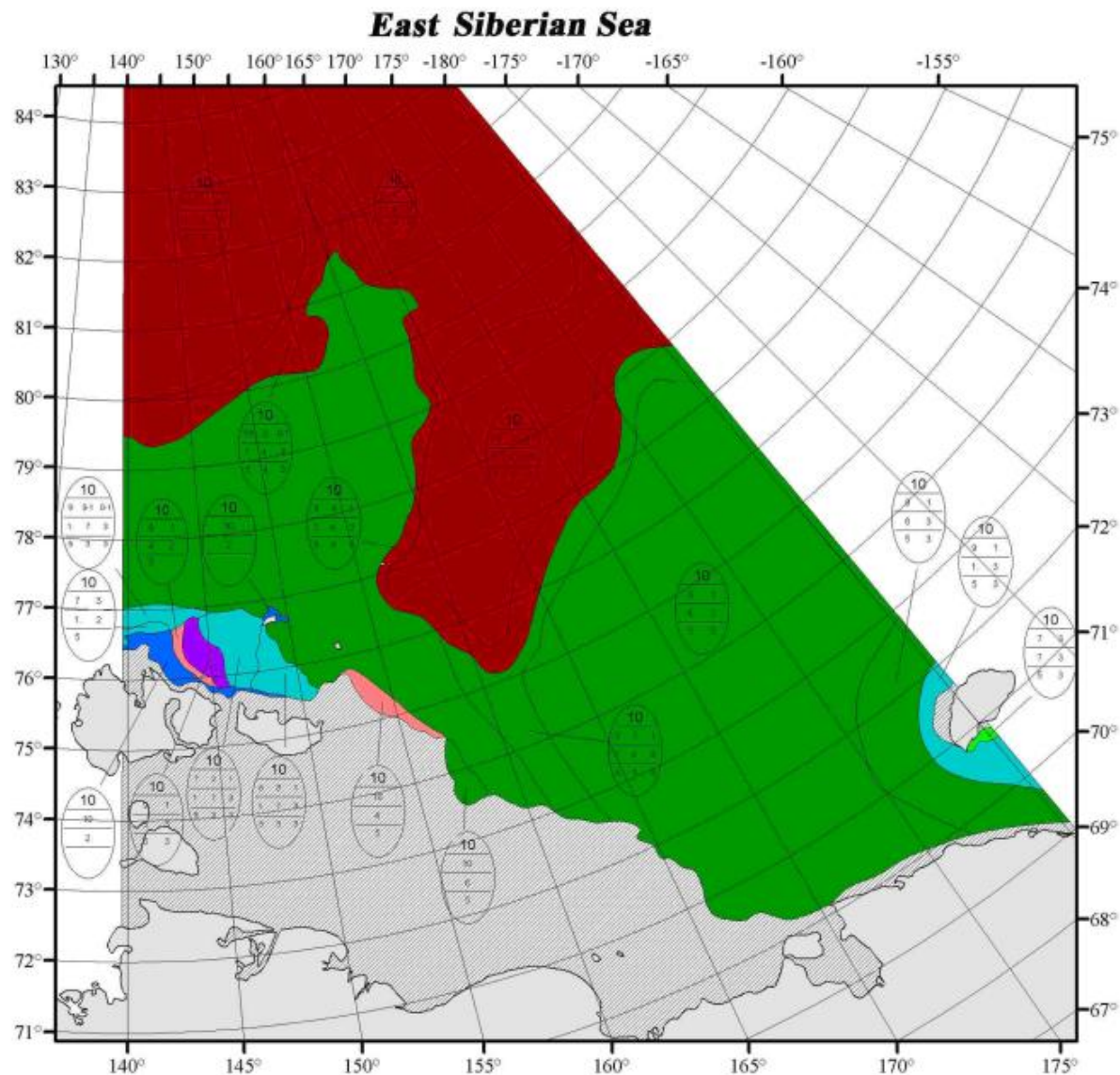
Relates to the “WMO Sea Ice Nomenclature”



Арктический и Антарктический  
научно-исследовательский  
институт Росгидромета

Arctic and Antarctic Research  
Institute of Roshydromet

**Feb. 28 - March. 02 2005**



Example sea ice chart with egg codes



# SSIII

Semantic Sea Ice Interoperability Initiative

[Home](#)

## Ontologies:

- *sea ice*
- *concentration*
- *development*
- *form*
- *ice of land origin*
- *Sigrid-3*
- *Egg Code*

## Ontology Browser

Load an Ontology: [Sea Ice](#) | [Seaice Concentration](#) | [Seaice Development](#) | [Seaice Form](#) | [Ice of Land Origin](#) | [Sigrid-3](#) | [Egg Code](#)

Ontology Browser v1.4.2 Help



All ontologies

Ontologies | Classes | Object Properties | Data Properties | Annotation Properties | Individuals | Datatypes | DL Query Options | Render labels ☒

### Classes

- + Thing
  - + **Forms of floating ice**
    - + Ice fragment
    - IceOfLandOrigin
  - + Ice with Development Stage
  - + SpatialObject

### Forms of floating ice permalink

<http://purl.org/wmo/seaice/form#IceWithForm>  

#### Annotations (3)

- source "[http://www.aari.nw.ru/gdsidb/docs/wmo/nomenclature/WMO\\_Nomenclature\\_draft\\_version1-0.pdf#Section1.1](http://www.aari.nw.ru/gdsidb/docs/wmo/nomenclature/WMO_Nomenclature_draft_version1-0.pdf#Section1.1)"
- comment "Any form of ice floating in water. The principal kinds of floating ice at the sea surface are sea ice which is formed by the freezing of sea water at the surface, lake ice and river ice formed on rivers or lakes and glacier ice (ice of land origin). The concept also includes ice that is grounded."
- label "Forms of floating ice"

#### Superclasses (1)

- Thing



# “Modeling the Model”

- Semantically representing Sea Ice Characteristics and the disposition of shortwave radiation in the community sea ice model (CICE) to depict:
  - Modeled processes (ice thickness change, melt pond formation, disposition of SW radiation)
  - The observations upon which model parameters are based
  - The observations that could be used to validate the model
- The goal is for the ontology to serve as a bridge between the operational, research and modeling communities looking at sea ice characteristics and polar climate interactions



Thank You  
[parsonsm@nsidc.org](mailto:parsonsm@nsidc.org)  
[nsidc.org/ssiii](https://nsidc.org/ssiii)



# DataONE (Observation Network for Earth): Enabling New Science by Supporting the Management of Data Throughout its Life Cycle



• Amber Budden, Roger Dahl, Rebecca Koskela, Bill Michener, Robert Nahf, Mark Servilla



• Dave Viegla



• Suzie Allard, Carol Tenopir, Maribeth Manoff, Robert Waltz, Bruce Wilson



• John Cobb, Bob Cook, Giri Palanisman, Line Pouchard



• Patricia Cruse, John Kunze



• Sky Bristol, Mike Frame, Richard Huffine, Viv Hutchison, Jeff Morisette, Jake Weltzin, Lisa Zolly



• Chad Berkley, Stephanie Hampton, Matt Jones



• Paul Allen, Rick Bonney, Steve Kelling



• Ryan Scherle, Todd Vision



• Randy Butler



• Ewa Deelman



• Peter Honeyman



• Jeff Horsburgh



• Robert Sandusky



• Bertram Ludaescher



• Peter Buneman



• Cliff Duke



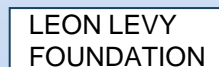
• Carole Goble



• Donald Hobern



• David DeRoure





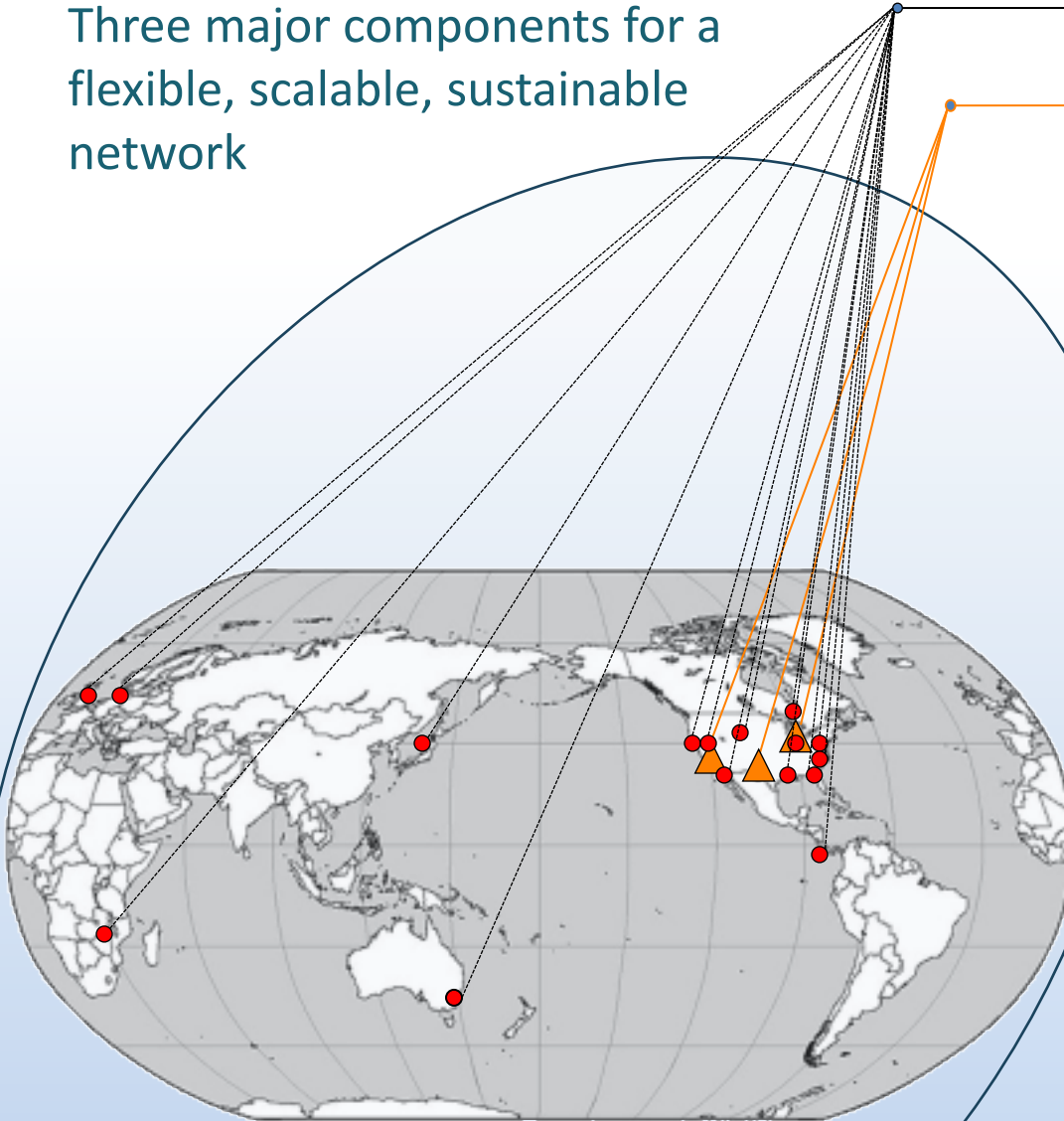
# DataONE Cyberinfrastructure

Three major components for a flexible, scalable, sustainable network

**Member Nodes**

**Coordinating Nodes**

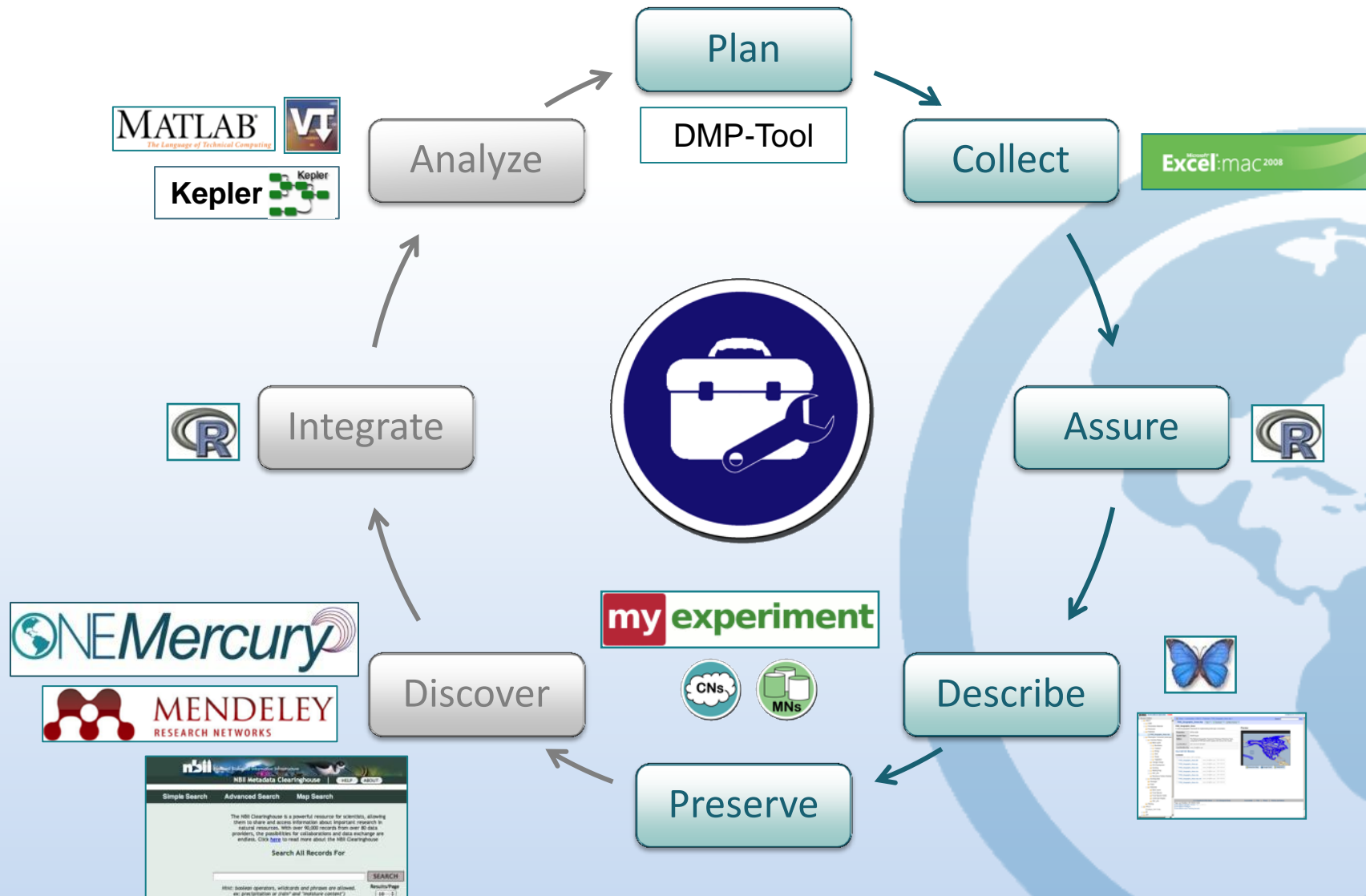
**Investigator Toolkit**



# The DataONE Federation




# Investigator Toolkit Activities



# Data Management Planning Tool

<http://dmp.cdlib.org>

November 1, 2011 = National Press Release




## DMP TOOL

Build your data management plan

[Contact Us](#) | [Sign up](#) | [Login](#)

[Home](#) | [About DMP Tool](#) | [DMP News](#) | [My Plans](#) | [Help](#)



Create ready to use data management plans for specific funding agencies

**The DMP Tool allows you to:**

Meet funder requirements for data management plans.

Get step by step instructions and guidance for your data management plan as you build it.

In many cases, get institution specific advice and assistance.

**Sign up and start building your data management plan now!**

**Data Management Plan**

As indicated in the past events and ongoing, the DMP Tool has a track record of helping staff, faculty or principal investigator of potential research data and how activity plans and communication the results with the scientific community in conferences and other scientific activities.

**Section 5: Types of data to be produced**

Types of data, research, research, software, software, software, software, and other materials to be produced in the course of the project.

The plan to manage and store research data produced under this award as well as the associated metadata that describes the data, including procedures, hardware, software and data analysis methods. Metadata, data or raw data, study of specific papers, plans for future studies in your research, communications with colleagues and other contacts are not included in this plan. Also included are data access, communication information, metadata necessary to be able to find data until they are archived, or any information protected under law.

**See a plan created with the DMP Tool**

**Recent DMP News**

[Open Access and Climate Research Data](#)

[Data, Data Everywhere...A Deluge of Data Management Articles](#)

[University of Illinois at Urbana-Champaign joins DMP Tool partners](#)

[Funder X now available in DMP Tool](#)

[more news >](#)

copyright 2011  
[Privacy statement](#) | [Terms of use](#)



# ONE Mercury

- An advanced data discovery tool
- Enables search and retrieval of content indexed by DataONE
- Primary web based user interface for DataONE
- Operates on each Coordinating Node
- Same SOLR / Lucene index is utilized by other client tools

The screenshot displays the ONE Mercury web interface. At the top, the logo "ONE Mercury" is shown. Below it, a search bar is labeled "Search For:" with a "Results/Page" dropdown set to "10" and a "SEARCH" button. A hint text reads: "Hint: boolean operators, wildcards and phrases are allowed. ex: precipitation or (rain\* and \"moisture content\")". Below the search bar are buttons for "Show/Hide Advanced Options" and "HELP".

The interface is divided into several sections:

- Fielded Search:** Contains three input fields for "FullText" and "OR" operators.
- Date Search:** Includes radio buttons for "Collection Date", "Publication Date", and "Either". It also has a "during" dropdown and date range inputs "mm/dd/yyyy" and "thru mm/dd/yyyy".
- Geographic Search:** Features a world map with labels for continents and oceans. To the right of the map are controls for "List Areas in:" (USA, WORLD), "Select from list", "Search Area:" (overlaps, encloses), and directional buttons (North, West, East, South). A "Place Name:" input field and a "view on map" button are at the bottom.
- Content Type:** A list box showing "All", "Maps and Data", "Publications", and "Tools and Software".
- Member Nodes:** A list box showing "All", "ORNL Distributed Active Archive Center for Biogeochemical Dynamics (OR)", "Dryad", "NBII Metadata Clearinghouse", "All NBII Partner Nodes", and "NBII Metadata Clearinghouse Principal Node".

At the bottom, a "Selected Query (Not Editable)" field is present.



Your search found: 10 documents.

Query: text : water

Filter by author	Filter by keywords	Filter by Originator
<a href="#">Cain James William (1)</a> <a href="#">Farnsworth Elizabeth (1)</a> <a href="#">Kratz Timothy (1)</a> <a href="#">Sandoval Cristina P. (1)</a> <a href="#">Van Schalkwyk (1)</a> <a href="#">Washburn Libe (1)</a> <a href="#">Wilson Gail (1)</a>	<a href="#">North Temperate Lakes - LTER (2)</a> <a href="#">abundance (2)</a> <a href="#">density (2)</a> <a href="#">plankton (2)</a> <a href="#">zooplankton (2)</a> <a href="#">2000 (1)</a> <a href="#">2005 (1)</a>	<a href="#">Channel Islands National Marine Sanctuary; Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO) (1)</a> <a href="#">Coal Oil Point Natural Reserve; University of California Natural Reserve System; University of</a>

[Prev](#) [Next](#)

[Return to Search](#)

Sort By:

**Relevance**

Date Range

Source

### Filter by Data Providers

[DEMO3 \(9\)](#)

[R2D2 \(1\)](#)

**ABIOTIC MONITORING OF PHYSICAL CHARACTERISTICS IN POREWATERS AND SURFACE WATERS OF MANGROVE FORESTS FROM THE SHARK RIVER SLOUGH AND TAYLOR SLOUGH, EVERGLADES NATIONAL PARK, SOUTH FLORIDA FROM DECEMBER 2000 TO MAY 2005** 01/01/36875 - 01/01/38498

Datasource: DEMO3

Data on porewater salinity, temperature, conductivity, pH and redox have been collected to help explain patterns found in porewater nutrient concentrations that were sampled in the same plots. ...

★★★★★★★★★

[Find similar data](#) [View full metadata](#)

**CINMS/PISCO: PHYSICAL OCEANOGRAPHY: BOTTOM-MOUNTED ADCP DATA: SAN MIGUEL ISLAND, CALIFORNIA, USA (BAYXXX)**

12/15/2005 - 03/23/2006

Datasource: R2D2

This metadata record describes bottom-mounted ADCP data collected at San Miguel Island, California, USA, by PISCO. Measurements were collected using an RDI 600 kHz Workhorse Sentinel ADCP beginning 2005-12-15. The instrument depth was 015 meters, in an overall water depth of 015 meters (both relative to Mean Sea Level, MSL). The instrument was programmed with a sampling interval of 2.0 minutes and a vertical resolution of 1 meter....

★★★★★★★★★

[Find similar data](#) [View full metadata](#)

**NORTH TEMPERATE LAKES LTER: ZOOPLANKTON - TROUT LAKE AREA**

04/22/1981 - 11/12/2008

Datasource: DEMO3

Zooplankton samples are collected from the seven primary northern lakes (Allequash, Big Muskellunge, Crystal, Sparkling, and Trout lakes and bog lakes 27-02 [Crystal Bog], and 12-15 [Trout Bog]) at two to nine depths using a 2 m long Schindler Patalas trap (53um mesh) and with vertical tows (1 m above the bottom of the lake to the surface) using a Wisconsin net (80um mesh). Zooplankton samples are preserved in buffered formalin (up until the year 2000) or 80% ethanol (2001 onwards) and archived. Data are summed over sex and stage and integrated volumetrically over the water column to p...

★★★★★★★★★

[Find similar data](#) [View full metadata](#)

**LITTLE ROCK LAKE EXPERIMENT: ZOOPLANKTON**

08/20/1983 - 10/23/2000

Datasource: DEMO3

The Little Rock Acidification Experiment was a joint project involving the USEPA (Duluth Lab), University of Minnesota-Twin Cities, University of

A Search Tool for Scientific Data
mercury.ornl.gov/pilotcatalog2/send/query?term1=water

Your search found: 1  
Query: text :

Filter by author

- Cain James William (1)
- Farnsworth Elizabeth (1)
- Kratz Timothy (1)
- Sandoval Cristina P. (1)
- Van Schalkwyk (1)
- Washburn Libe (1)
- Wilson Gail (1)

Filter by keyword

- North Temperate Lakes LTER (2)
- abundance (2)
- density (2)
- plankton (2)
- zooplankton (2)
- 2000 (1)

- Coastal Oceans (PISCO) (1)
- Coal Oil Point Natural Reserve; University of California Natural Reserve System; University of

Select which items you'd like to add to your library

- ☐ Abiotic monitoring of physical characteristics in porewaters a...
- ☒ CINMS/PISCO: Physical Oceanography: bottom-mounted ADC...
- ☐ North Temperate Lakes LTER: Zooplankton - Trout Lake Area
- ☐ Little Rock Lake Experiment: Zooplankton
- ☒ Behavior, Activity Patterns and Foraging Strategies of Beaver (...)
- ☒ Soil Warming Experiment - Phenology and Growth of Vegetat...
- ☐ Bone density and Calcium and Phosphorus content of the gir...
- ☐ Nest Success of the Yellow Warblers (Dendroica petechia) and...
- ☐ Mycorrhizal Suppression Plots
- ☐ Coal Oil Point Natural Reserve Annual Report

Select All
Deselect All

Cancel
OK

Prev 1 Next

Search Mercury Pilotcatalog

Sort By:

Relevance
Date Range
Source

Filter by Data Providers

- DEMO3 (9)
- R2D2 (1)

**ABIOTIC MONITORING OF PHYSICAL CHARACTERISTICS IN POREWATERS AND SURFACE WATERS OF MANGROVE FORESTS FROM THE SHARK RIVER SLOUGH AND TAYLOR SLOUGH, EVERGLADES NATIONAL PARK, SOUTH FLORIDA FROM DECEMBER 2000 TO MAY 2005**

01/01/36875 - 01/01/38498

*Datasource:* DEMO3

Data on porewater salinity, temperature, conductivity, pH and redox have been collected to help explain patterns found in porewater nutrient concentrations that were sampled in the same plots. ...

★★★★★★★★★★

Find similar data
View full metadata

**CINMS/PISCO: PHYSICAL OCEANOGRAPHY: BOTTOM-MOUNTED ADCP DATA: SAN MIGUEL ISLAND, CALIFORNIA, USA (BAYXXX)**

12/15/2005 - 03/23/2006

*Datasource:* R2D2

This metadata record describes bottom-mounted ADCP data collected at San Miguel Island, California, USA, by PISCO. Measurements were collected using an RDI 600 kHz Workhorse Sentinel ADCP beginning 2005-12-15. The instrument depth was 015 meters, in an overall water depth of 015 meters (both relative to Mean Sea Level, MSL). The instrument was programmed with a sampling interval of 2.0 minutes and a vertical resolution of 1 meter....

★★★★★★★★★★

Find similar data
View full metadata

**NORTH TEMPERATE LAKES LTER: ZOOPLANKTON - TROUT LAKE AREA**

04/22/1981 - 11/12/2008



Email Query Bookmark Query RSS Feed for Query Help

Your search found: 10 documents.  
Query: text : water

Filter by author	Filter by keywords	Filter by Originator
<a href="#">Cain James William (1)</a> <a href="#">Farnsworth Elizabeth (1)</a> <a href="#">Kratz Timothy (1)</a> <a href="#">Sandoval Cristina P. (1)</a> <a href="#">Van Schalkwyk (1)</a> <a href="#">Washburn Libe (1)</a> <a href="#">Wilson Gail (1)</a>	<a href="#">North Temperate Lakes - LTER (2)</a> <a href="#">abundance (2)</a> <a href="#">density (2)</a> <a href="#">plankton (2)</a> <a href="#">zooplankton (2)</a> <a href="#">2000 (1)</a>	<a href="#">Channel Islands National Marine Sanctuary; Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO) (1)</a> <a href="#">Coal Oil Point Natural Reserve; University of California Natural Reserve System; University of</a>

Prev 1 Next

Search Mercury Pilotcatalog

Sort By: Relevance Date Range Source

Filter by Data Providers

[DEMO3 \(9\)](#)  
[R2D2 \(1\)](#)

ABIOTIC MONITORING OF PHYSICAL CHARACTERISTICS IN POREWATERS AND SURFACE WATERS OF MANGROVE FORESTS FROM THE SHARK RIVER SLOUGH AND TAYLOR SLOUGH, EVERGLADES NATIONAL PARK, SOUTH FLORIDA FROM DECEMBER 2000 TO MAY 2005

01/01/36875 - 01/01/38498

Datasource: DEMO3

Data on porewater salinity, temperature, conductivity, pH and redox have been collected to help explain patterns found in porewater nutrient concentrations that were sampled in the same plots. ...

★★★★★★★★

Find similar data View full metadata

My Library

- Presentations
- Unfiled Items
- Trash

Group Libraries

- biodiversity, standards

Display all tags in this library

0 tags selected Deselect All

Title	Creator	
Behavior, Activity Patterns and Foraging Strategie...	Sagehen Creek Fiel...	
CINMS/PISCO: Physical Oceanography: bottom-m...	Channel Islands Na...	
Content Standard for Digital Geospatial Metadata ...		
Darwin Core	Darwin Core Task ...	1
Ecological Metadata Language (EML) Specification		
Soil Warming Experiment - Phenology and Growt...		

Info Notes Tags Related

Item Type: Web Page

Title: Content Standard for Digital Geospatial Metadata — Federal Geographic Data Committee

Author: (last), (first)

Abstract:

Website Title:

Website Type:

Date:

Short Title:

URL: http://www.fgdc.gov/metadata/csdg...

Accessed: Sat May 7 07:20:37 2011

Language:

Rights:

# DataONE Users Group

- represents the **needs and interests** of the stakeholders
- provides **feedback** on DataONE services
- **reviews** DataONE policies and procedures
- identifies and promotes **sustainability** approaches
- maintains effective **communication** and coordination





# DataONE.org







New DataONE Education Resources



Coming Soon

Latest News | es Live Posted: 10/7/2011 DataONE Learning Modules available online Posted: 08/22/2011

About	Participate	Products	Education	Data
What is DataONE? DataONE Organization	DataONE Users Group Member Nodes	OneMercury Investigator Toolkit	Training Activities Education Modules	Find Contribute





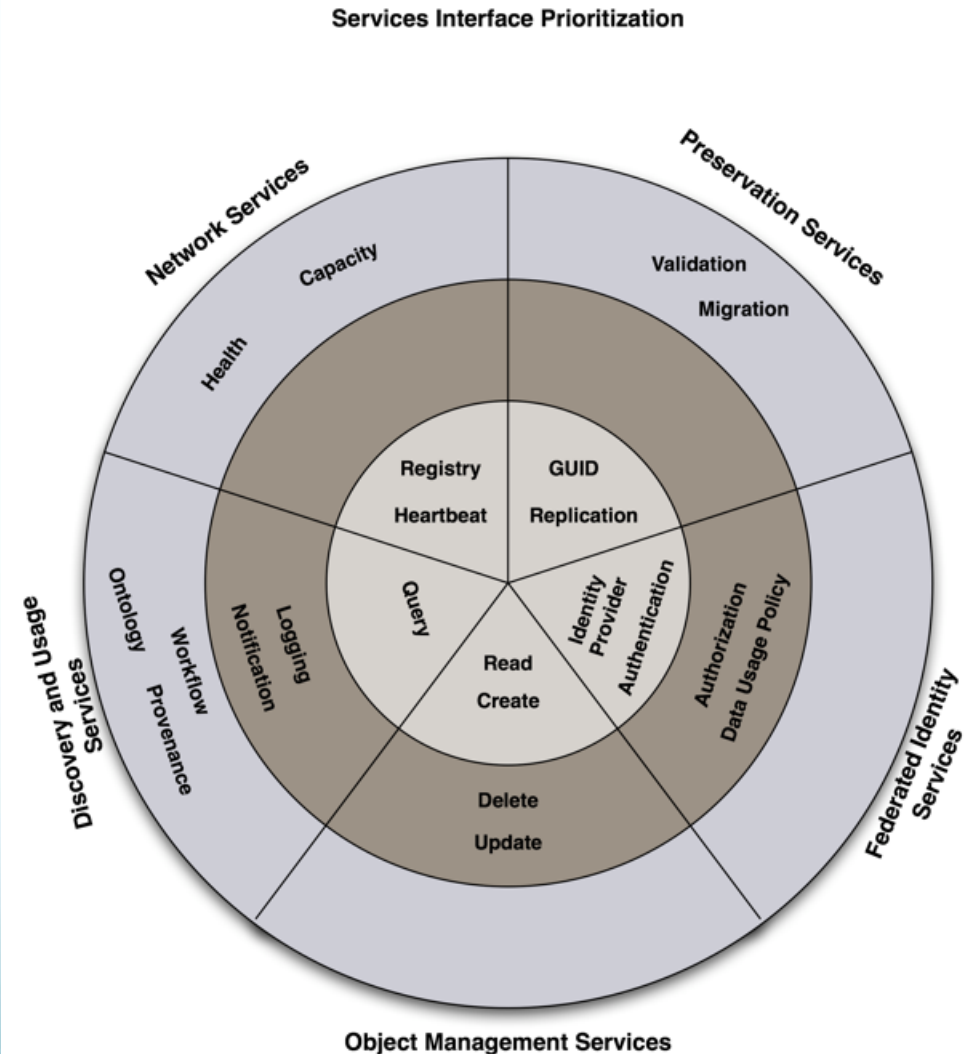
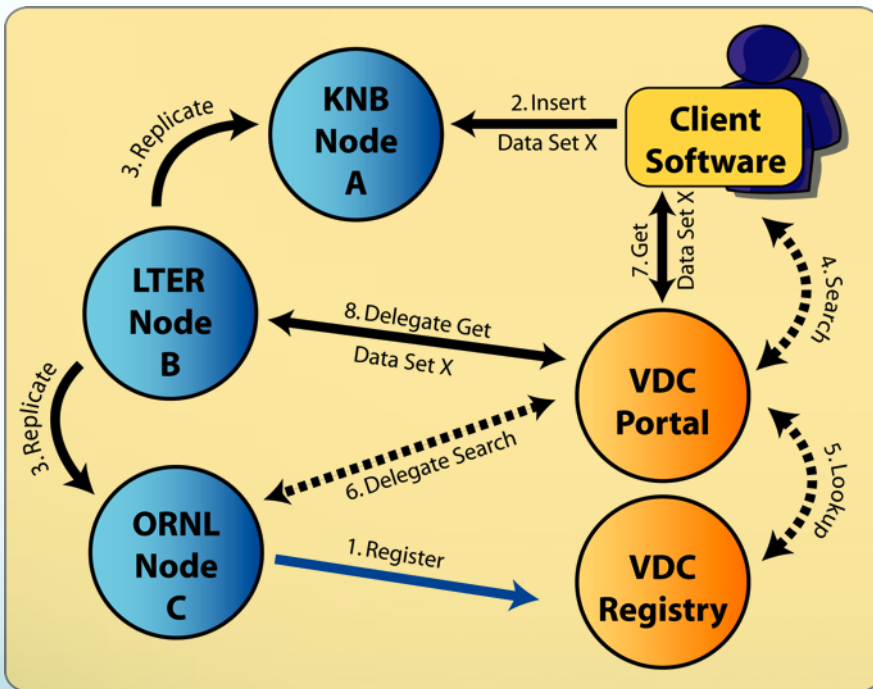
# INTEROP: Creation of a Virtual Data Center for the Biodiversity, Ecological and Environmental Sciences

- **William Michener, Mark Servilla** – University of New Mexico
- **Matthew Jones** – National Center for Ecological Analysis and Synthesis, UC-Santa Barbara
- **Kathleen Smith, Hilmar Lapp, Todd Vision** – National Evolutionary Synthesis Center (NESCent), Duke University and university of North Carolina
- **Robert Cook**, Oak Ridge National Laboratory DAAC for Biogeochemical Dynamics
- **Mike Frame**, USGS National Biological Information Infrastructure
- **Dave Vieglais**, University of Kansas

# Objective 1 (Technical Working Group):

- The goal of this project is to **develop new community capacity and new technologies to support design, implementation, and deployment of a Virtual Data Center (VDC)** for biodiversity, ecological and environmental data.

# Architecture Definition and Services Interface Prioritization



Each ring of the diagram represents decreasing priority as one moves out from the center of the diagram.

# Completion of 38 Use Case Scenarios

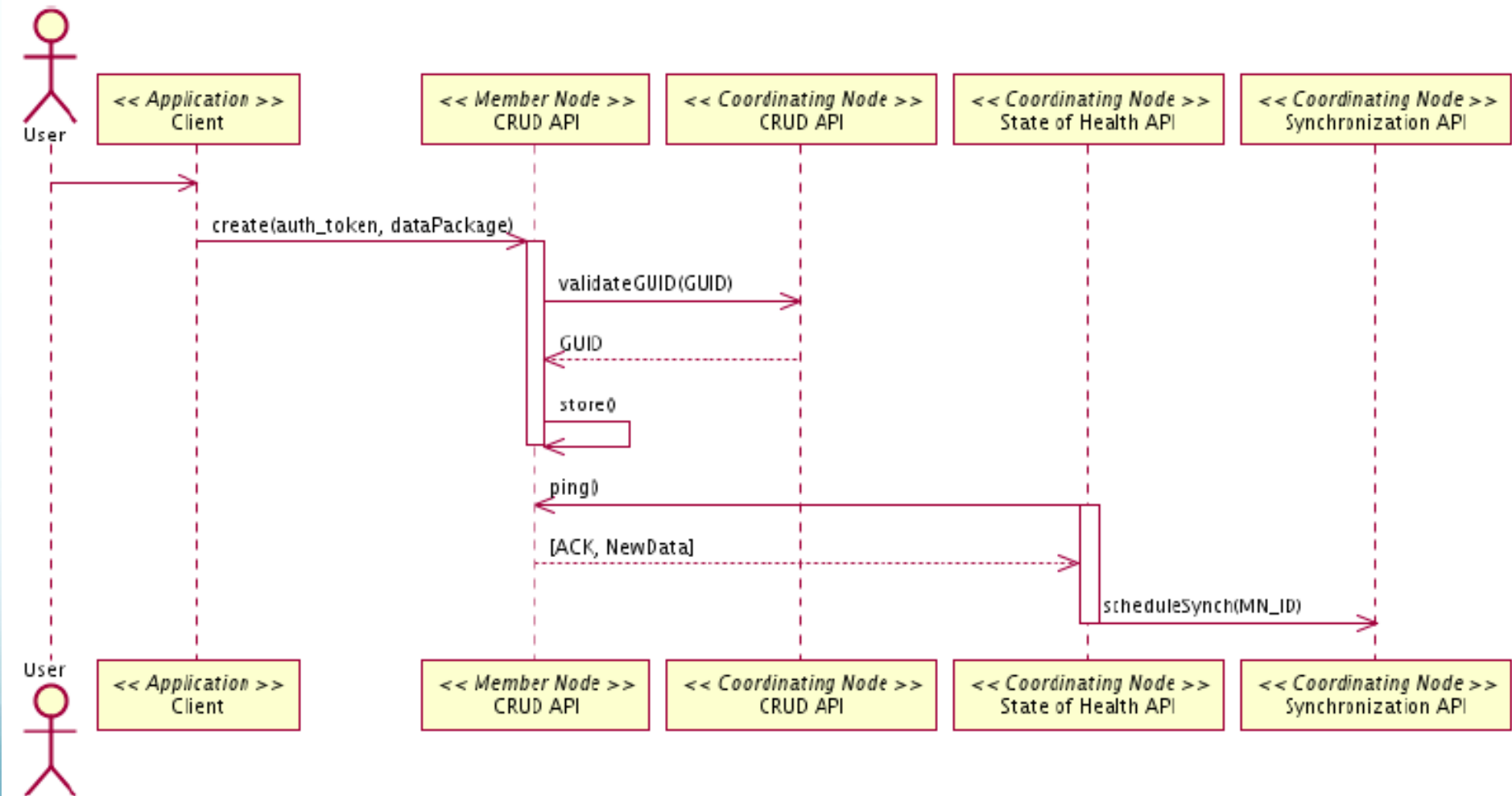


Figure 5.8: Create, update, delete, search metadata or data object in Member Node.

# VDC Prototyping Activities

1. The `d1_common_python` library which provides low level implementation of the Virtual Data Center service interfaces.
2. The Command Line Client which forms a part of an Investigator Toolkit that can be used by scientists to manage and analyze data that are part of the VDC infrastructure.
3. The Generic Member Node (GMN) which is a standalone data repository that implements service interfaces designed under the VDC project.
4. The prototype Dryad Member Node which extended the GMN to enable VDC services on the Dryad repository.
5. The prototype ORNL DAAC Member Node which extended the GMN to enable VDC services on the ORNL DAAC repository.



# Objective 2: Community Engagement Working Group

- Address **socio-cultural barriers** to data preservation and data sharing, as well as **data center and VDC sustainability**.
- Workshops
  - Research Coordination Workshop (INTEROPs +)
  - Federated Security Workshop
  - Virtual Data Center Preservation Strategy Workshop
  - Virtual Data Center Stakeholders Group Meeting
  - Community Best Practices Workshops
  - Data Governance Workshop



### Best Practices

Search Best Practices

Contains  [SEARCH](#)

Best Practice Categories

[All Best Practices](#)

- Content and Structure (15)
- Data Access and Discovery (5)
- Data Documentation (3)
- Data Preservation and Archives (7)
- Planning Policies and Governance (1)
- Quality Assurance and Quality Control (5)
- Vocabulary Standards and Services (2)

Featured Best Practice

[Define the contents of Data Files](#)

Category: Data Documentation

- Formats for dates, time, geographic coordinates, and other parameters
- Define any coded values
- Quality flags or qualifying values
- Define missing values

### Search Tools

Contains  [SEARCH](#)

Tool Categories

[All Tools](#)

- Analysis & Modeling (34)
- Data Acquisition & Modeling (30)
- Workflow (7)

Featured Tool

**S-PLUS (S+)**

Primary Category: Analysis & Modeling

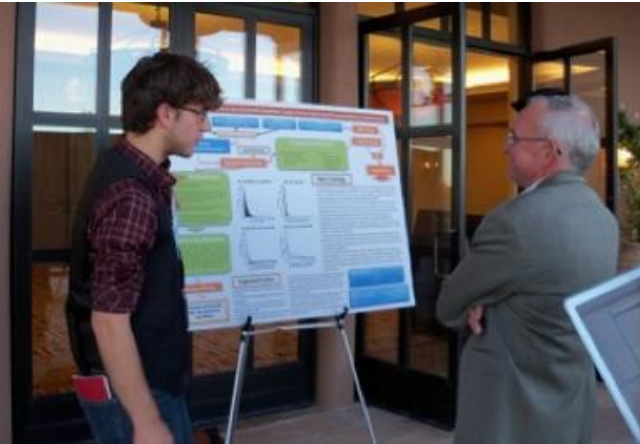
S-PLUS is a commercial implementation of the S statistical programming language that was developed by Bell Labs. S+ has a cross-platform integrated development environment (IDE), provides the ability to analyze gigabyte class data sets on the desktop, and a package system for deployment of analytics.

Cost: Cost-basis

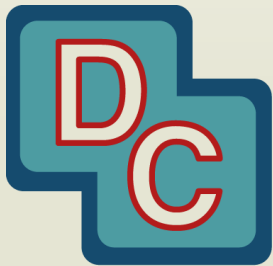
**TIBCO Spotfire S+**

# Objectives 3, 4 & 5: Education, Outreach and Training

- Four + graduate students participated annually in a **summer cyberinfrastructure traineeship** that is modeled after the Google Summer of Code.



- **Outreach will be provided at annual meetings** of relevant professional societies, emerging environmental observatories, and research networks.
- Project results will be incorporated in numerous **annual training sessions** supported by project partners and through the broad adoption of the interoperability solutions developed as part of this project.
  - Outreach and Training
    - American Geophysical Union
    - Ecological Society of America
    - Society of Fish and Wildlife Information Managers
    - Dozens of Professional Society meetings



# Data Conservancy

---

Sayeed Choudhury

NSF DataNet/Interop Pls meeting – January 26, 2012



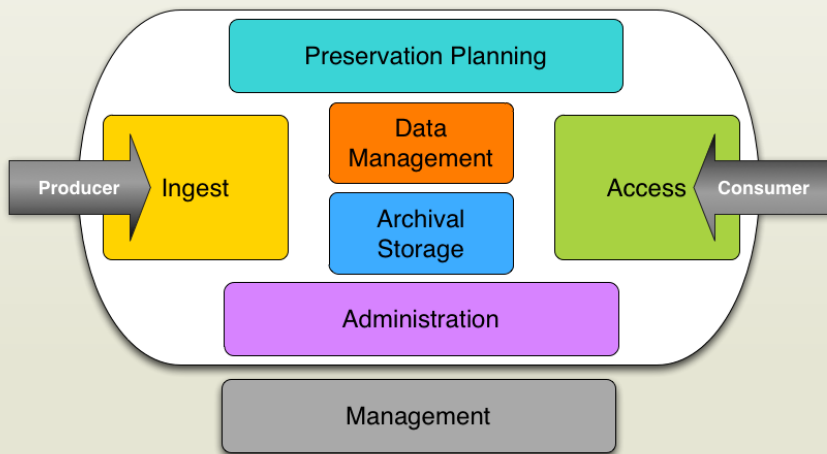
# Data Conservancy Objectives

---

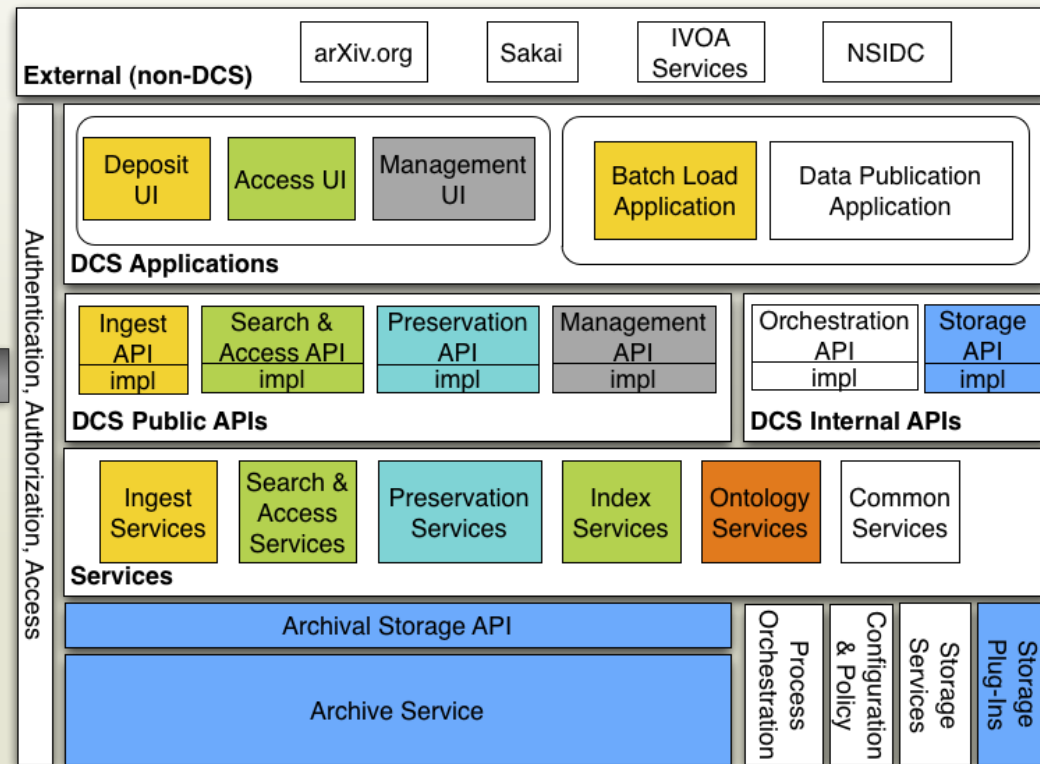
- Data Conservancy is a community that develops solutions for data preservation and sharing to promote cross-disciplinary re-use.
- Preserve – collect and take care of research data
- Share – reveal data's potential and possibilities
- Discover – promote re-use and new combinations



# Architecture mapped to OAIS



Open Archival Information System  
Functional Entities



Data Conservancy Service  
Architecture Block Diagram





# Definition of Data Preservation

---

- “Data preservation involves providing enough representation information, context, metadata, fixity, etc. such that someone other than the original data producer can use and interpret the data.”
  - Ruth Duerr, National Snow and Ice Data Center



# Defining Sustainability



- “Ensuring that valuable digital assets will be available for future use is not simply a matter of finding sufficient funds. It is about mobilizing resources—human, technical, and financial—across a spectrum of stakeholders diffuse over both space and time.”



## Powered by Data Conservancy

---

- JHU Data Management Service (DMS) represents the culmination of two years of research, design, development and implementation of Data Conservancy
- Service launched in July 2011
- DC instance launched in October 2011
- Important, essential foundations in place
- There remains work to be done



# Establishing the JHU DMS

---

- May 2010 NSF announces DMP expectations
- Services incubated and scoped summer/fall 2010
  - Build on Data Conservancy expertise
- Proposed in January and launched in July 2011
  - Consultative data management planning services to support NSF proposals
  - Post award data management services
- Assessment of service in March 2012



# Acknowledgements and Resources

---

- NSF Award OCI-o830976
- Sheridan Libraries financial support
- Johns Hopkins University financial support
- Elliot Metsger for infrastructure slide
- Data Conservancy colleagues for their exceptional work and patience
- <http://dataconservancy.org>
- <http://dmp.data.jhu.edu>



# A Feasibility Study for an NSF Open-Access Publication Repository

Sayed Choudhury

NSF DataNet/Interop Pls meeting

January 26, 2012

# Feasibility Study

- NOT an implementation plan (i.e., not a decision by NSF but rather a “what if...” analysis)
- Johns Hopkins University, University of Michigan and Council on Library and Information Resources
- Advantages and disadvantages of different approaches based on framework developed through earlier study funded by Open Society Institute
- Technical, policy and business dimensions

# Workshops

- Convened three workshops to focus on each of these dimensions
- Participants included a diverse group of potential stakeholders, service providers, partners from both non-profit and for-profit sectors
- Feedback from reports resulted in ideas about scope and potential technical approaches

# Scope

- Consensus support for connection between publications and data
- PI acts as proxy for “publication” for public release (typically those identified within Fastlane interim or final reports)
- Does not include administrative reports or “grey” literature (e.g., Powerpoint files)
- Overwhelming response that NSF needs to define specifications or requirements

# Four Categories of Technical Approaches

- Category 1 – Locally installed system (e.g., ePrints, DSpace)
- Category 2 – Large, scale hosted systems (e.g., HathiTrust, Pubmed Central)
- Category 3 – Federation of systems (i.e., content distributed and harvested)
- Category 4 – Custom solution
- Final report by end of February 2012



# Acknowledgements

- NSF OCI/Interop award # 948134
- Mark Cyzyk, Tim DiLauro (Johns Hopkins)
- John Wilkin, Jeremy York (Michigan)
- Amy Friedlander (formerly CLIR; now NSF SBE)
- Workshop participants
- ContentDM and Digital Commons who completed framework questions



# DRInet in 5 Minutes

- an NSF Interop project

Purdue University



# Developing Community-based DRought Information Network Protocols and Tools for Multidisciplinary Regional-Scale Applications (DRInet)

- NSF funded Data Interoperability project
- Started in Jan. 2009
- Multidisciplinary team
  - Carol Song (HPC)
  - Daniel Aliaga (CS)
  - Jake Carlson (Library)
  - Indrajeet Chaubey (ABE)
  - Rao Govindaraju (CE)
  - Chris Hoffmann (CS)
  - Dev Niyogi (Ag/EAS)
  - Lan Zhao (HPC)
  - Grad and undergraduate students, data expert





# Objective & Challenges

- Implement and explore *computational infrastructure* to create a science base for drought information collection, fusion of heterogeneous data relevant to droughts, and for facilitating the broadest possible participation of the community
- Challenges
  - Various information is out there, but not standardized, not utilized to their potential.
  - Regional and local level – lacking systematic way of collecting information and data, data synthesis and dissemination
  - Ad hoc data collection mechanism
  - Lack of ways to include uncertainty information in drought classification
  - Heterogeneous dataset
  - Diverse needs for data and information



# Approaches

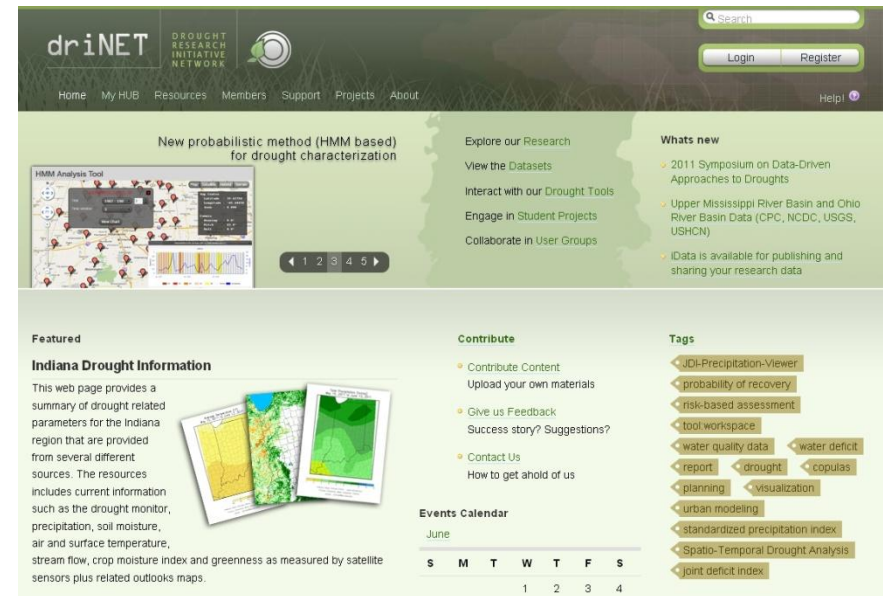
- Community driven
  - Stakeholders on advisory board
  - Workshop in 2011 (participation from multiple domains, state and national operations centers and agencies, industry, students, international)
  - Data/info needs from diverse user base
- Research areas
  - Drought classification
  - Drought implications on water quality
  - Drought implications on air quality
  - Data synthesis
  - Visualization-based decision tools
  - Metadata (balance of depth & breadth, standards & local needs)
- Collaborate with other projects
  - U2U (USDA)
  - Teaching





# Cyberinfrastructure

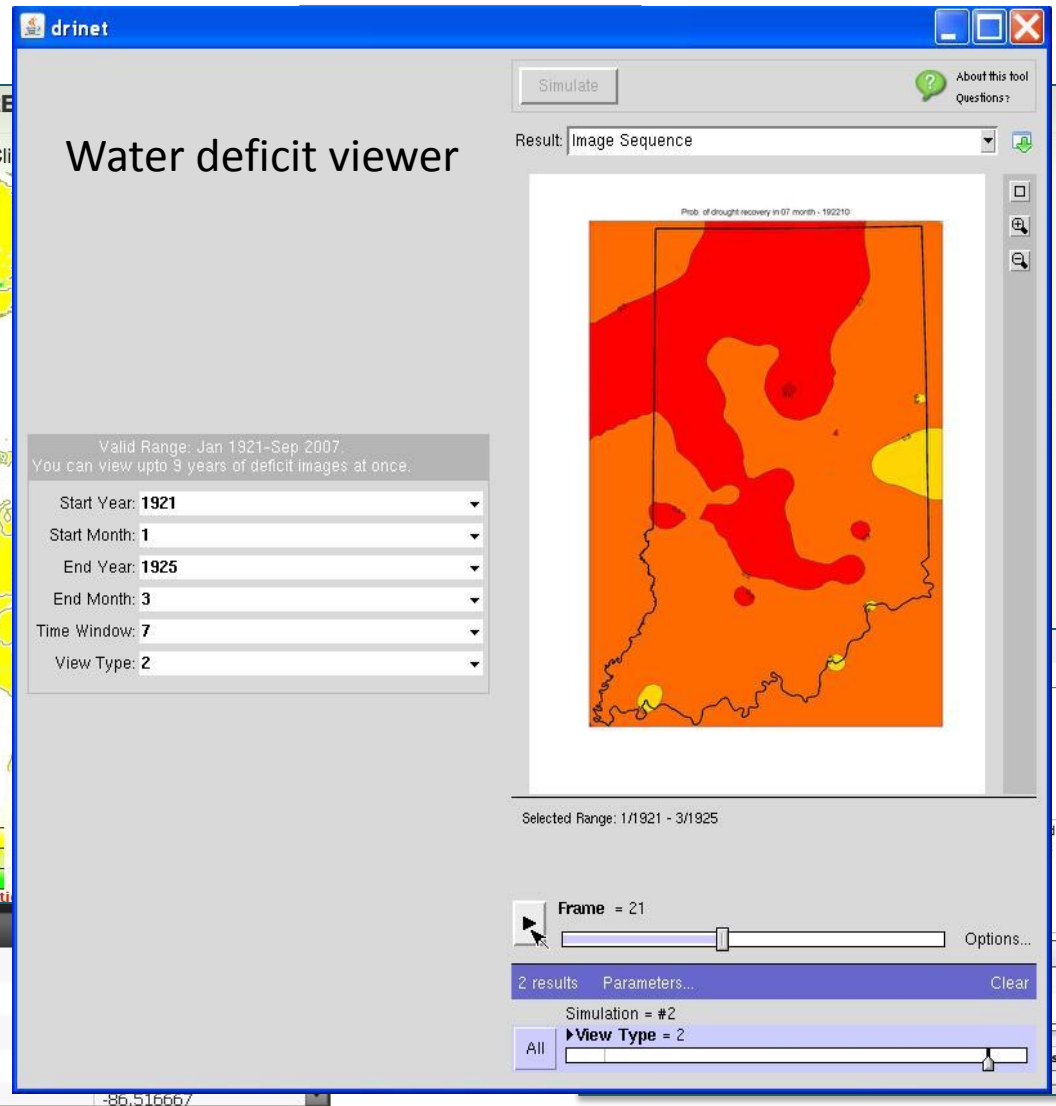
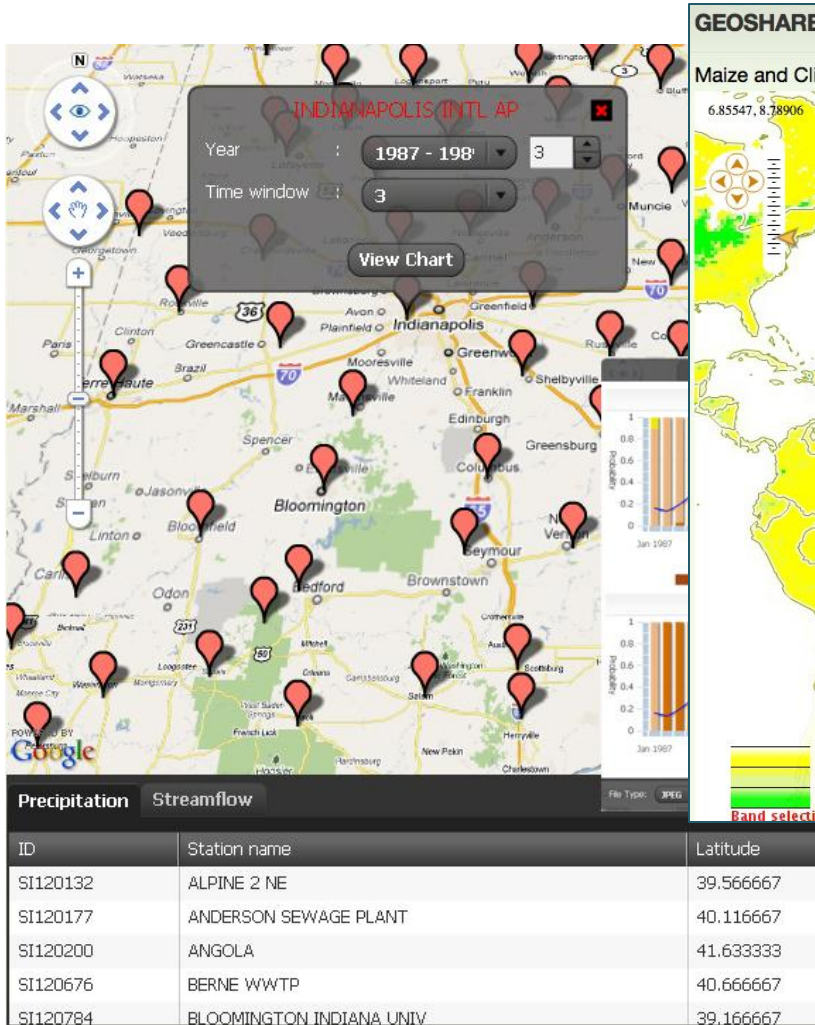
- DRI-net CI based on HUBzero for
  - Online interactive tools
  - Sharing and user participation
  - One stop shop for researchers, students and end users
- Expanded capabilities
  - Distributed data access
  - Self data and metadata publishing
  - Map-based navigation
  - Geospatial data enabled
  - Connecting data with tools, models, visualization





# Interactive Tools

## HMM-based Probabilistic Drought Classification





# Data

## Water quality data

### Matson Ditch AT DeKalb CR 39, IN

- Variable: Bacteria
- Sub Var: Ecoli
- Original Num of Data Points: 221
- Start Date: 1996/09/25
- End Date: 2008/10/29

Show Details Get Data Plot

### Monthly Average of Ecoli Count at Matson Ditch AT DeKalb CR 39, IN



## Historical Averages of Precipitation and Temperature Data

This dataset include plots of historical time series of average precipitation and temperature data, courtesy of the Oklahoma State University.

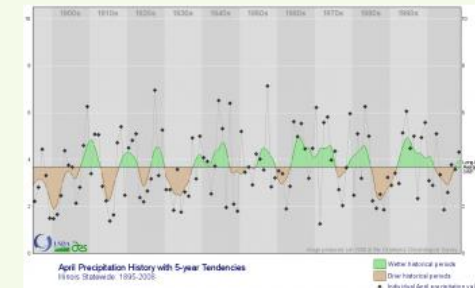
IL

- IL.zip
- trace.IL-CD00.prcp.Annual.png
- trace.IL-CD00.prcp.April.png
- trace.IL-CD00.prcp.August.png
- trace.IL-CD00.prcp.Autumn.png
- trace.IL-CD00.prcp.December.png
- trace.IL-CD00.prcp.February.png
- trace.IL-CD00.prcp.January.png
- trace.IL-CD00.prcp.July.png
- trace.IL-CD00.prcp.June.png

File Name:

trace.IL-CD00.prcp.April.png

Preview:

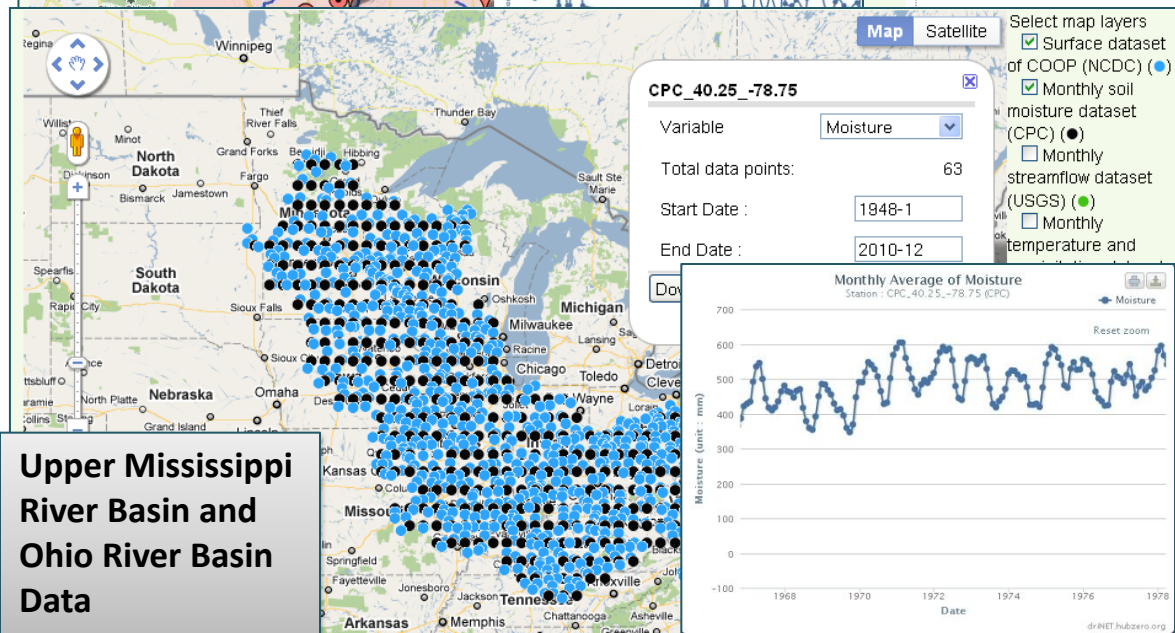
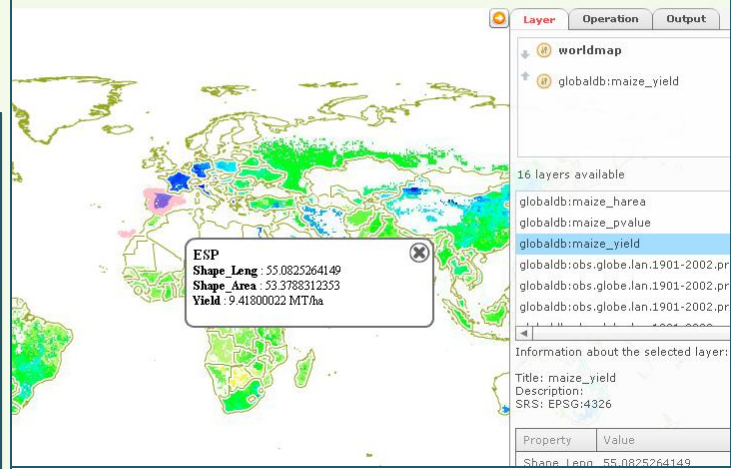


View graph in full size

File size ( 22758 Byte)

## yield data

(Double click on available layers to browse the data, and click on Operation to see a preview)



### CPC\_40.25\_-78.75

Variable: Moisture

Total data points: 63

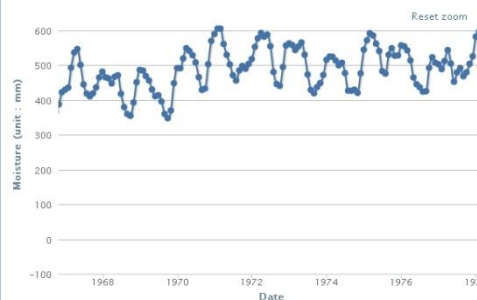
Start Date: 1948-1

End Date: 2010-12

Download

### Monthly Average of Moisture

Station: CPC\_40.25\_-78.75 (CPC)







# Data Publishing

**iDATA** Publish, Browse & Discover

Data explorer

Current Path : MASSIE CREEK

	Date/Time	Level	Temp	S
MASSIE CREEK				
• Date/Time				
• Level				
• Temp				
• Sp Cond				
• pH				
• Turb+				
• ODOsat				
• ODO				

<input type="checkbox"/>	2010-10-01 22:00	0.324	14.97	8
<input type="checkbox"/>	2010-10-01 22:30	0.322	14.85	8
<input type="checkbox"/>	2010-10-01 23:00	0.321	14.74	8
<input type="checkbox"/>	2010-10-01 23:30	0.32	14.64	8
<input type="checkbox"/>	2010-10-02 00:00	0.32	14.51	8
<input type="checkbox"/>	2010-10-02 00:30	0.319	14.41	8
<input type="checkbox"/>	2010-10-02 01:00	0.319	14.31	8
<input type="checkbox"/>	2010-10-02 01:30	0.318	14.21	8

## Sharing Massie Creek Water Quality Dataset

Select a group to share the collection with

Group

drinetteam

Privilege

Read + Append

Shared with:

Group Name	Access	Date
public	Read only	2011/03/19
drinetteam	Read+Append	2011/03/19

## Tabular Data Import Wizard (Step 1 of 4)

- Schema Definition
- Input File
- Delimiter
- Verification
- Execution

### Define a database schema

Specify a schema for the tabular data

Name	Data Type	UQ
Datetime	Datetime	<input type="checkbox"/>
Avg Level (m)	Real number	<input type="checkbox"/>

## Tabular Data Import Wizard

- Schema Definition
- Input File
- Delimiter
- Verification
- Execution

### Importing data

Now, importing data into database. Please wait...

Importing data
Creating a database table...
The database table was successfully created.
Uploading the file ...
Upload completed.
Importing data into database...
Importing completed.



Finish



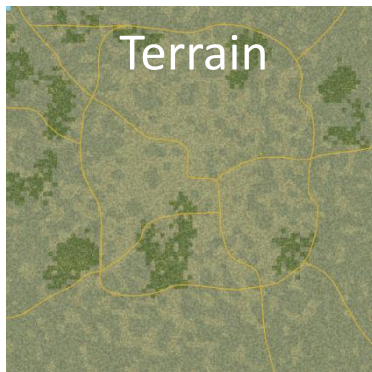
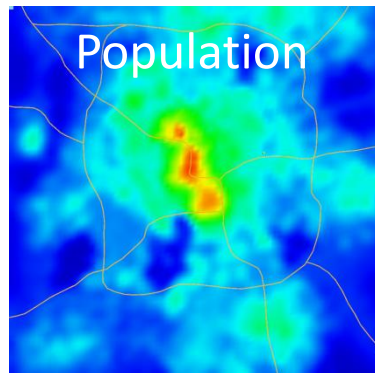
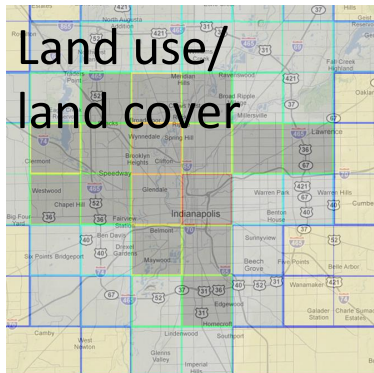
Close



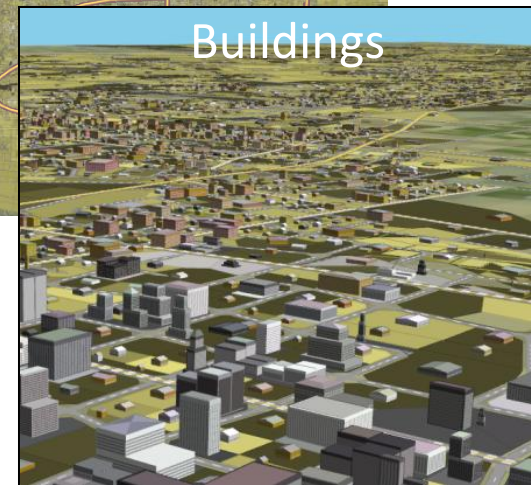
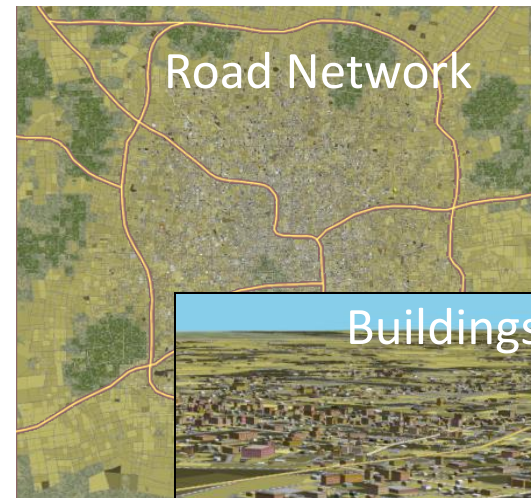
# Urban Weather Modeling



Data



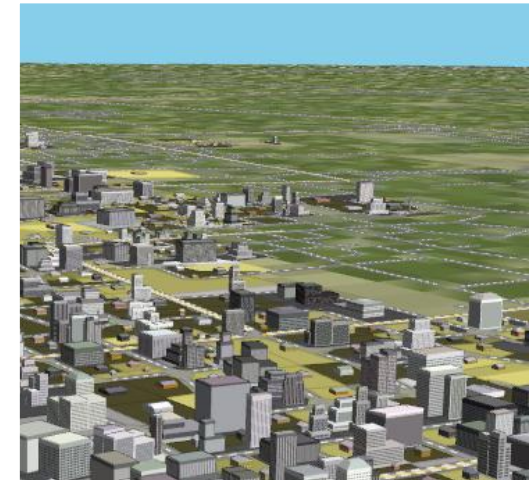
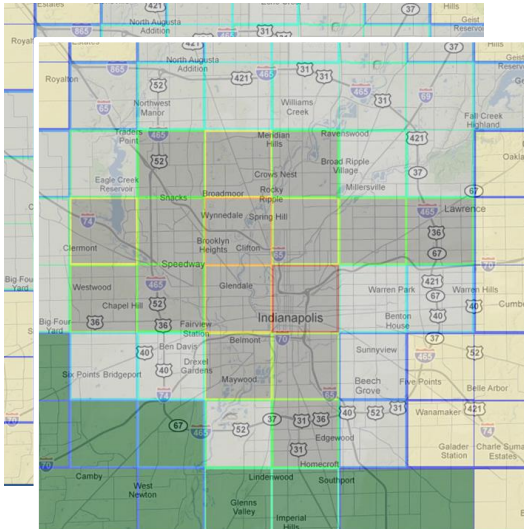
Result







# Urban Weather Modeling



Modify land use (e.g.,  
application of greening policy)

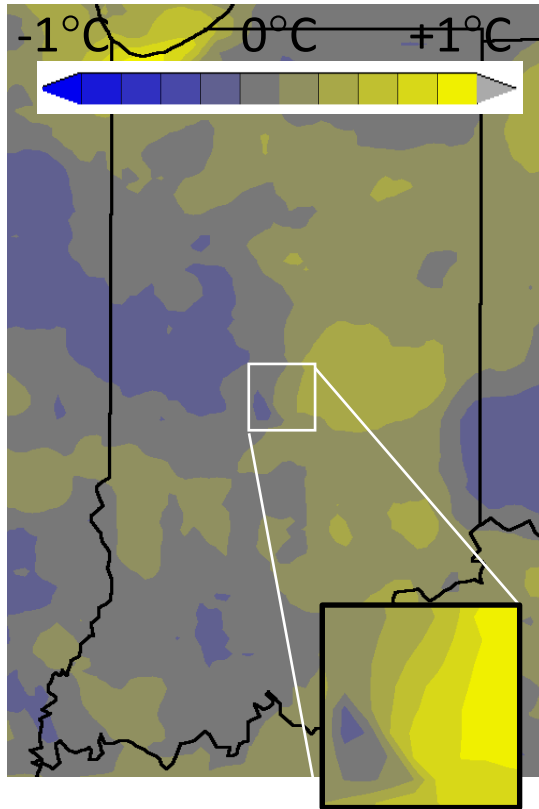
New fully instantiated city model is produced



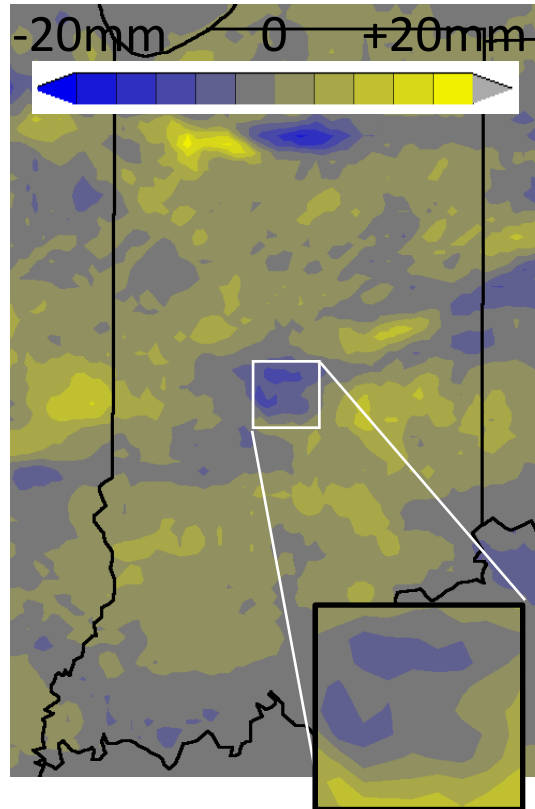
# Example Changes in Local Weather



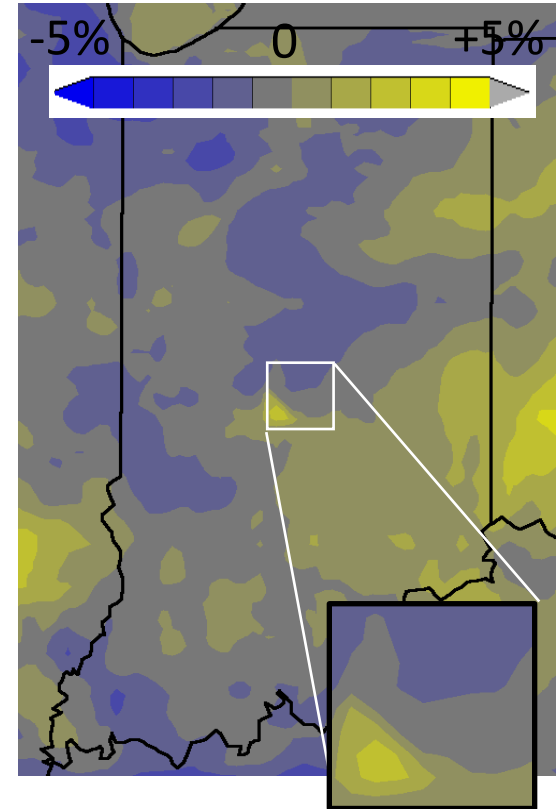
Temperature



Rainfall



Humidity



# Instabilities and Trust: Basic Research Data Under Construction

Sharon Traweek

Gender Studies & History Departments

University of California, Los Angeles

Data 2012, January 26, 2012

# Funding, 2009-2014

- \* The Transformation of Knowledge, Culture, and Practice in Data-Driven Science: A Knowledge Infrastructures Perspective, Award# 20113194, Alfred P. Sloan Foundation, 2012-2014
- \* EAGER: Knowledge and Data Transfer: the Formation of a New Workforce, Award# 1145888, National Science Foundation, September 2011-August 2012
- \* The Data Conservancy: A Digital Research and Curation Virtual Organization, Award# 0830976, National Science Foundation, August 2009-July 2012

# Collaborators

- Christine Borgman, PI, UC Presidential Chair in Information Studies, UCLA

<http://is.gseis.ucla.edu/cborgman/> <http://works.bepress.com/borgman/>

- Sharon Traweek, co-PI, Gender Studies & History, UCLA

<http://www.history.ucla.edu/traweek/>

[http://www.womensstudies.ucla.edu/faculty\\_traweek.html](http://www.womensstudies.ucla.edu/faculty_traweek.html)

- Laura Wynholds and Ashley Sands, Graduate Student Researchers, Information Studies, UCLA

<http://is.gseis.ucla.edu/academics/degrees/phd/students.htm>

- Former member: David Fearon, Entrepreneurial Library Program, Johns Hopkins University

<http://www.library.jhu.edu/elp/Team/fearon.html>



# Knowledge Infrastructures in Astronomy Data, 2009-2014

<http://knowledgeinfrastructures.gseis.ucla.edu/index.html>

Research Questions:

What are the current practices in data design, collection, access, use, reuse, revision, sharing, and curation for basic research in astronomy?

What are the variations in those practices?

Who is doing this work?

How is all this changing and why ?

# Our Research Methods

- Interviews and oral histories
- Multi-sited ethnographies at universities, laboratories, and conferences
- Following databases: formation, sharing access, revisions, curation
- Following data from design and collection to publication and citation.

Our multi-sited ethnographic & oral history research includes large-scale database-driven astronomy projects (such as SDSS & LSST)

1. Large international, distributed teams (>100)
2. Huge archived digital databases from sky surveys
3. Data used to generate images for research
4. Increased access to research data
5. Increased workforce gathering & interpreting data
6. Changing organizational and funding infrastructure
7. Support by government/industry/university sectors
8. Two primary examples: SDSS & LSST

# Basic Research

- State-of-the-art project design, equipment, data collection, databases, and analytic tools are not standardized and always under repair.
- Stable equipment, data, and analytic tools are used to calibrate backgrounds against which new kinds of data can be recognized and the long process of data evaluation can proceed.
- Basic research communities know how to build knowledge in the context of this mixture of stability and instability in design, equipment, data, and analysis. They develop different approaches to answering shared questions.

# Sloan Digital Sky Survey (SDSS)

<http://www.sdss.org/>

<http://www.sdss.org/collaboration/>

- Sloan I 2000–05; Sloan II 2005–08; Sloan III 2008-14
- Funded by Sloan Foundation
- Telescope at Apache Point, NM, USA
- Partnership with Microsoft: Skyserver Database

<http://skyserver.sdss.org/dr1/en/skyserver/>

- \* SDSS database access rates: 952,336 monthly 2009-11 <http://skyserver.sdss.org/log/en/traffic/>



# Large Synoptic Survey Telescope

<http://www.lsst.org>

- LSST is the next big project in astronomy
- Scheduled to be ready by 2014
- Telescope at Cerro Pachón, Chile
- Collaborations now forming
- Includes many physicists
- Partnership with Google to download raw data on internet every three nights

# One astronomer's comment on access to LSST data

“LSST is the first large facility ... where the data [will be] public in thirty seconds. Every high school student in the world... [will have] the same access to the data that I do.... That means, if you're interested in something, you can get the data.”

# Variation in Astronomy Research Practices

Types of Observatories:

Ground & Space

Data Deposit Sites:

Facilities & Desktops

Investigators based at:

Universities,

Observatories, &

Space Research Facilities

Funding sources:

Govt agencies

Private foundations

Topic & Forms of Inquiry:

Extra-terrestrial signals

Star formation

NASA archives

Exoplanets

Galaxy evolution

X-ray space telescopes

Gamma-ray bursts

**Radio waves**

**Microwave &  
Infrared**

**Ultraviolet**

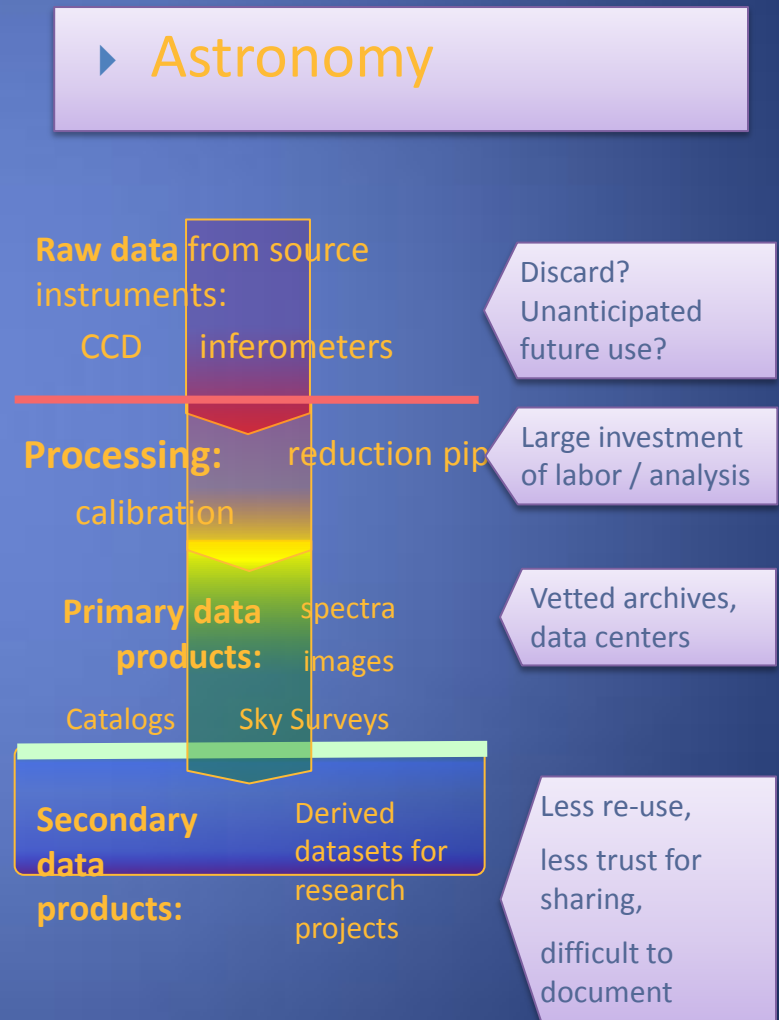
**X rays**

**Gamma rays**

Sky Surveys

# Findings: Building Trust in Data

- Building trust in **primary data**
  - Iterative calibration & testing
  - Vetted data systems and services
  - Human expertise
- Building trust in **secondary data**
  - Difficult to discover
  - Inconsistent documentation
  - N.I.H.



# Findings: stable & unstable data coexist in basic research

Data curation for basic research must accommodate the on-going daily work in basic research with both stable and instable data (data undergoing the long process of evaluation), including those in large-scale data sets.



# National Science Foundation (NSF)EAGER Program

- EArly-concept Grants for Exploratory Research  
...support exploratory work in its early stages on untested, but potentially transformative, research ideas or approaches. NSF is in the process of implementing a new emphasis on transformative research ...

[http://www.nsf.gov/pubs/policydocs/pappguide/nsf09\\_29/gp\\_g\\_2.jsp#IID2](http://www.nsf.gov/pubs/policydocs/pappguide/nsf09_29/gp_g_2.jsp#IID2)

# Women and Minority Astronomers Strategic Engagement with Distributed, Multi-Disciplinary Collaborations and Large Scale Databases

- An NSF OCI funded research project 2009-2012
- The goal is to understand how gender, ethnicity, class, and nationality in the astronomy workforce intersects with changing knowledge-making practices.

# Collaborators

- PI: Sharon Traweek, UCLA, anthropologist & historian of 20c physical sciences (Europe, Japan, & US)
- Co-PI: Jarita Holbrook, University of Arizona, PhD, Astronomy & Astrophysics. Ethnographer of cultural astronomy practices (Fiji, Tunisia, UK, & US)
- Postdoctoral Research Associate: Reynal Guillen, UCLA Chicano Studies Research Center; BS/MS in astronomy & space physics & PhD, history of science. Ethnographic historian of minorities in science & engineering in southwestern US

### 3 Graduate Student Research Assistants

- Diane Yu Gu (from China), Doctoral Candidate, Comparative Higher Education, UCLA: mentoring women grad students in physical sciences
- Luis Felipe R. Murillo (from Brazil), Graduate Student, Cultural Anthropology, UCLA: networks among open source groups in Brazil, Japan, & US
- Brad Fidler (from Canada), UCLA PhD, big pharma marketing strategies with big data

# **CUNY High-performance Computing Initiative**

Paul Muzio

Nikolaos Trikoupis



- The CUNY HPC Center acknowledges support for the following:
- “Andy”, a SGI cluster with NVIDIA GPUs, was funded by:
  - NSF Grant 0855217
  - A grant from the NYC Council through the efforts of Councilman James Oddo
- “Salk”, a Cray XE6, was funded by:
  - NSF Grant 0958379
- MRI for Data Storage, was funded by:
  - NSF Grant 1126113
- New York State Regional Economic Development Grant

- Support CUNY's "Decade of Science" Initiative and the Integrated University Concept of Operation.
- Support the University's research and educational activities by making state-of-the-art HPC resources and expert technical assistance available to faculty and students.
- With CUNY faculty and researchers, develop proposals for external funding.
- Support National, New York State, and New York City initiatives in economic development.
- Support National and New York State initiatives to promote the sharing of HPC resources and technical knowledge.
- Support educational outreach programs designed to encourage intermediary and high school students to pursue higher education and careers in science and technology.

- Senior Colleges
  - City College
  - Hunter College
  - Baruch College
  - Brooklyn College
  - Queens College
  - New York City College of Technology
  - College of Staten Island
  - John Jay College of Criminal Justice
  - York College
  - Lehman College
  - Medgar Evers College
- Community Colleges
  - Bronx Community College
  - Queensborough Community College
  - Borough of Manhattan Community College
  - Kingsborough Community College
  - LaGuardia Community College
  - Hostos Community College
- Graduate and Professional Schools
  - CUNY Graduate Center
  - Sophie Davis School of Biomedical Education
  - School of Law
  - William E. Macaulay Honors College
  - Graduate School of Journalism
  - School of Professional Studies
- Largest urban university in the United States.
  - 260,000 students in degree credit programs
  - 250,000 adult, continuing and professional education students.
  - College Now
  - Academic enrichment program for 32,500 high school students offered at CUNY campuses and more than 300 high schools.
  - University Teacher Academy
  - Free tuition for highly motivated mathematics and science majors who commit to teach in NYC
  - Extensive STEM and LSAMP Programs
- Demographics:
  - 61% female, 39% male
  - 0.2% American Indian/Native Alaskan, 15.8% Asian/Pacific Islander, 27.1% Afro-American, 25.7% Hispanic, 31.2% White
  - 68% attended New York City public high schools
  - 42% first time college
  - A total of 195 different languages are spoken
- Alumni
  - 12 Nobel Laureates
  - 2 Field Medal Awardees
  - Numerous recipients of Pulitzer Awards
  - 4 Rhodes scholars in last 6 years
  - Andy Grove, Robert Kahn, Jonas Salk

# Geographical Dispersion





# WHERE ARE WE?

- College of Staten Island
- 4500 square feet (computer room and offices)
- UPS and diesel backup









1. “Andy” was funded under NSF Grant 0855217 and the NYC Council (James Oddo)
2. “Salk” was funded under NSF Grant 0958379

# Existing Systems (Jan 2012)

System	Cores	Chip	Memory/ core (GB)	Disk (Tbyte)	Interconnect
Athena	344	Woodcrest	2	3.2	Ethernet
Zeus	64	Harpertown	2	7.8	Ethernet
	16		2		Ethernet
Bob	232	Barcelona	2	7.1	Infiniband
Andy	744	Nehalem	3	36.0 (Lustre)	Dual and Quad rail Infiniband
	96 x 448	GPU - FERMI	3		PClex Gen2
Salk	1,280	Magny-Cours	2	126.0 (Lustre)	Custom
Karle	24	Westmere	4 / 96	Lustre	SMP

Bob – Robert Kahn, co-developer TCP/IP

Andy – Andy Grove, co-founder, Intel

Salk – Jonas Salk, developer of the polio vaccine

Karle – Jerome Karle, mathematician, chemist, Nobel Laureate

- Linux
- PBSpro
- Lustre file system
- Cray compilers
- CUDA
- GNU compilers
- Intel compilers
- PGI compilers
- PGI Accelerator Programming Model
- OpenMPI
- Subversion
- AMD AMCL
- Atlas
- Cray LIBSCI
- FFTW
- GSL
- GMF
- IMSL
- Intel MKL
- Matlab
- Mathematica
- R
- SAS
- Sparsekit
- Stata

- Bamova (Bayesian Analysis of Molecular Variance)
- Bayescan
- Beast
- BEST
- BPP2
- GenomePOP2
- IMa2
- LAMARC
- Migrate
- MrBayes
- MSMS
- RAxML
- Structurama
- Structure



- Dalton
- DL-Poly
- Gaussian03
- Gaussian09
- Gromacs
- Hondo Plus
- HOOMD
- LAMMPS
- NAMD
- NWchem
- Octopus

- ADCIRC Coastal Circulation and Storm Surge Model
  - ARC\_solar\_radiation
  - Community Earth System Model (CESM)
  - Finite Volume Community Ocean Model (FVCOM)
  - Regional Ocean Modeling System (ROMS)
  - Weather Research Forecasting System (WRF)
  - WRF-chem
  - Model Evaluation Tools
    - PB2NC
    - ASCII2NC
    - PCP Combine tool
    - NCEP
    - NCAR
    - GRADS
    - Gen Poly Mask Tool
    - Grid Stat Tool
    - Mode
    - Wavelet Stat Tool
    - Stat Analysis
    - Mode Analysis
  - NetCDF, PnetCDF, HDF4, HDF5
- 
- Network Simulator 2
  - Phoenix
  - Transims

- NSF Grant awarded in August 2011
- Storage Area Network
- Remote tape silo for backup and archiving
  - Fiber optic connection
- System is in pre-procurement stage
- Evaluating iRODS installation to support research activities and promote data sharing
  - Hosting a meeting with RENCI on 7 February 2012

- To support the activities of the 7 major research programs (and others)
  - Ecology, Evolutionary Biology, and Behavior Subprogram at the CUNY Graduate Center
  - Environmental CrossRoads Initiative
  - NOAA Cooperative Remote Sensing Science and Technology Institute
  - CUNY Institute for Sustainable Cities
  - Center for Maritime Systems, Stevens Institute of Technology
  - Institute for Macromolecular Assemblies
  - Subotnick Financial Services Center – Baruch College

Thank you

Questions?





Center for  
Technology in Government

# I-Choose: Building Information Sharing Networks to Support Consumer Choice

**Holly Jarman, Ph.D.**

[hjarman@albany.edu](mailto:hjarman@albany.edu)

Department of Public Administration and Policy,  
Rockefeller College of Public Affairs and Policy  
University at Albany, SUNY

# Who are we?



# The Project

- Coffee grown in Mexico, consumed in Canada & USA
- Research Tasks:
  - Ontology development
  - Data architecture
  - Stakeholder network evaluation
  - Policy recommendations
- Broad evidence base: case studies (Mexico, Canada, Guatemala), interviews, certification databases, stakeholder focus groups, policy & legal review

# Which Coffee Would You Buy?



\$9.58



\$2.49

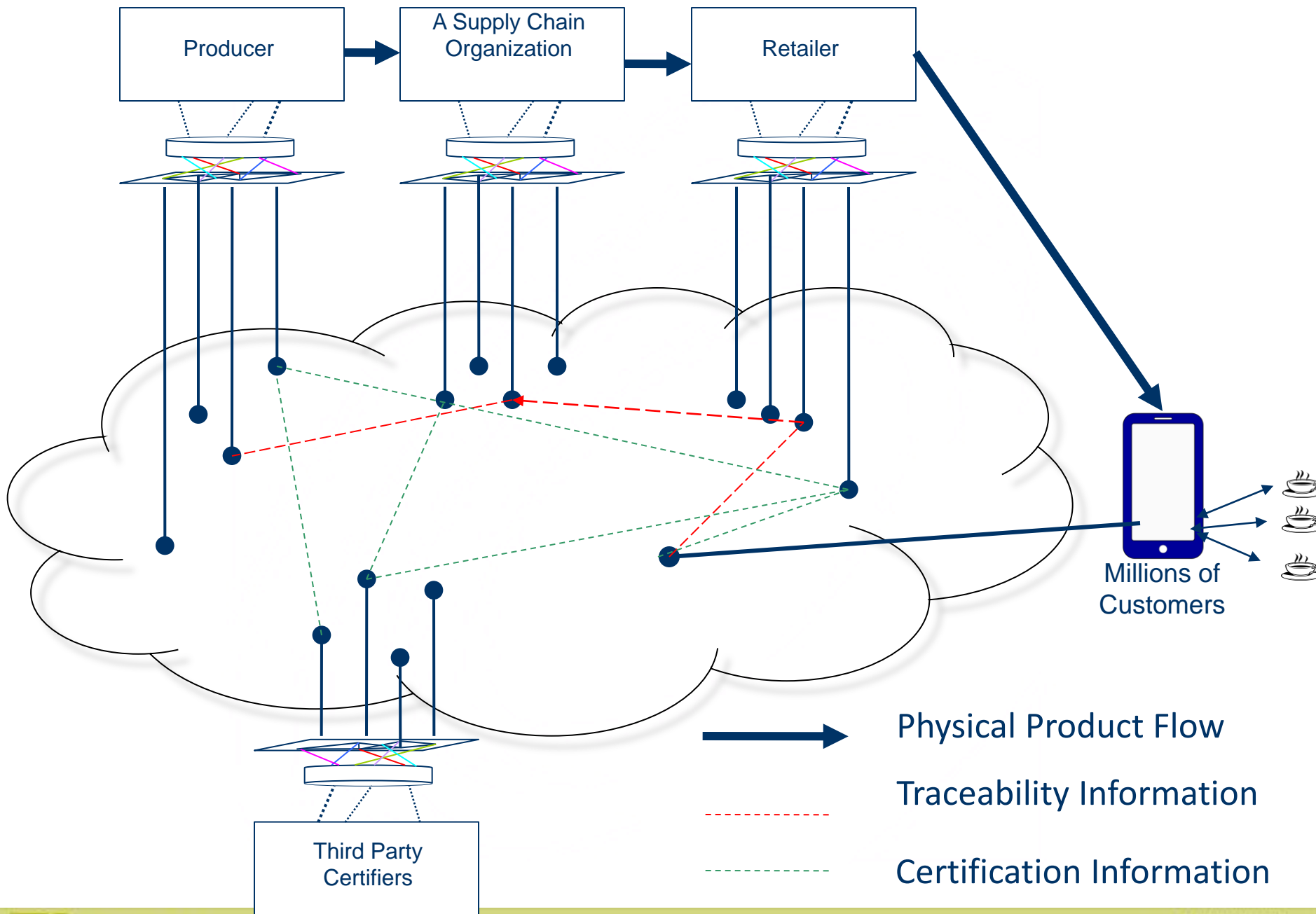


\$8.00



\$9.34

(All prices normalized to US\$ per pound)

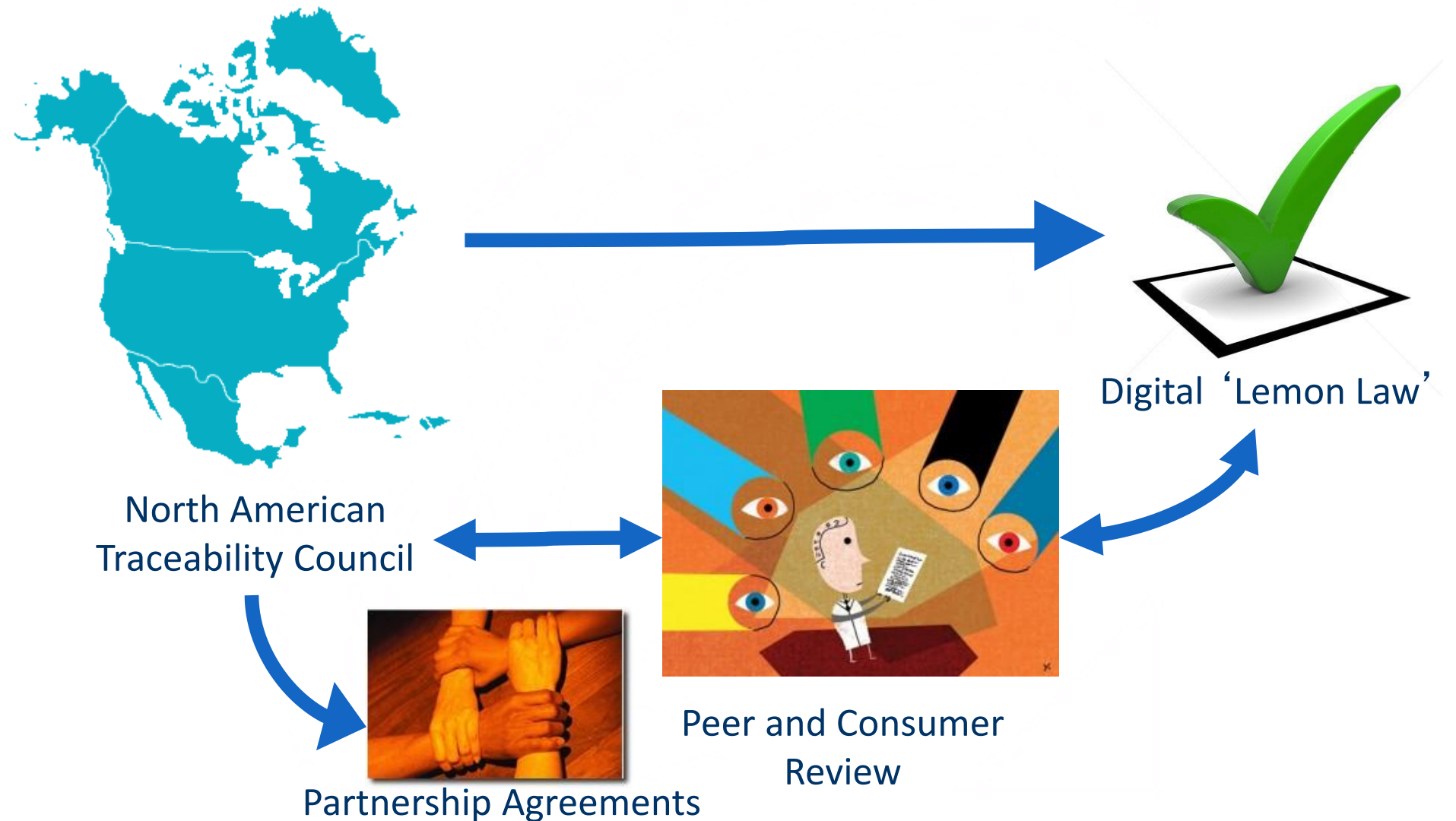




# Governance: Towards Private Sector Transparency

- Understanding and Modeling Certification and Inspection Regimes
- State vs. Non-State Led Regulatory Systems
- Balancing Cost and Sustainability
- Balancing Privacy and Access to Information

# Governing I-Choose



# Thank You



**The following slide-set has  
unpublished content removed.**



# General Data Analysis and Visualization using HPC

The University of Tennessee Center for Remote Data Analysis & Visualization (RDAV)

---

Jian Huang  
Associate Professor | EECS  
Associate Director | RDAV  
University of Tennessee



Funded by NSF Teragrid XD, RDAV is a joint effort by University of Tennessee, National Center for Supercomputing Applications (NCSA), Oak Ridge National Laboratory, Lawrence Berkeley National Laboratory and University of Wisconsin.



Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation



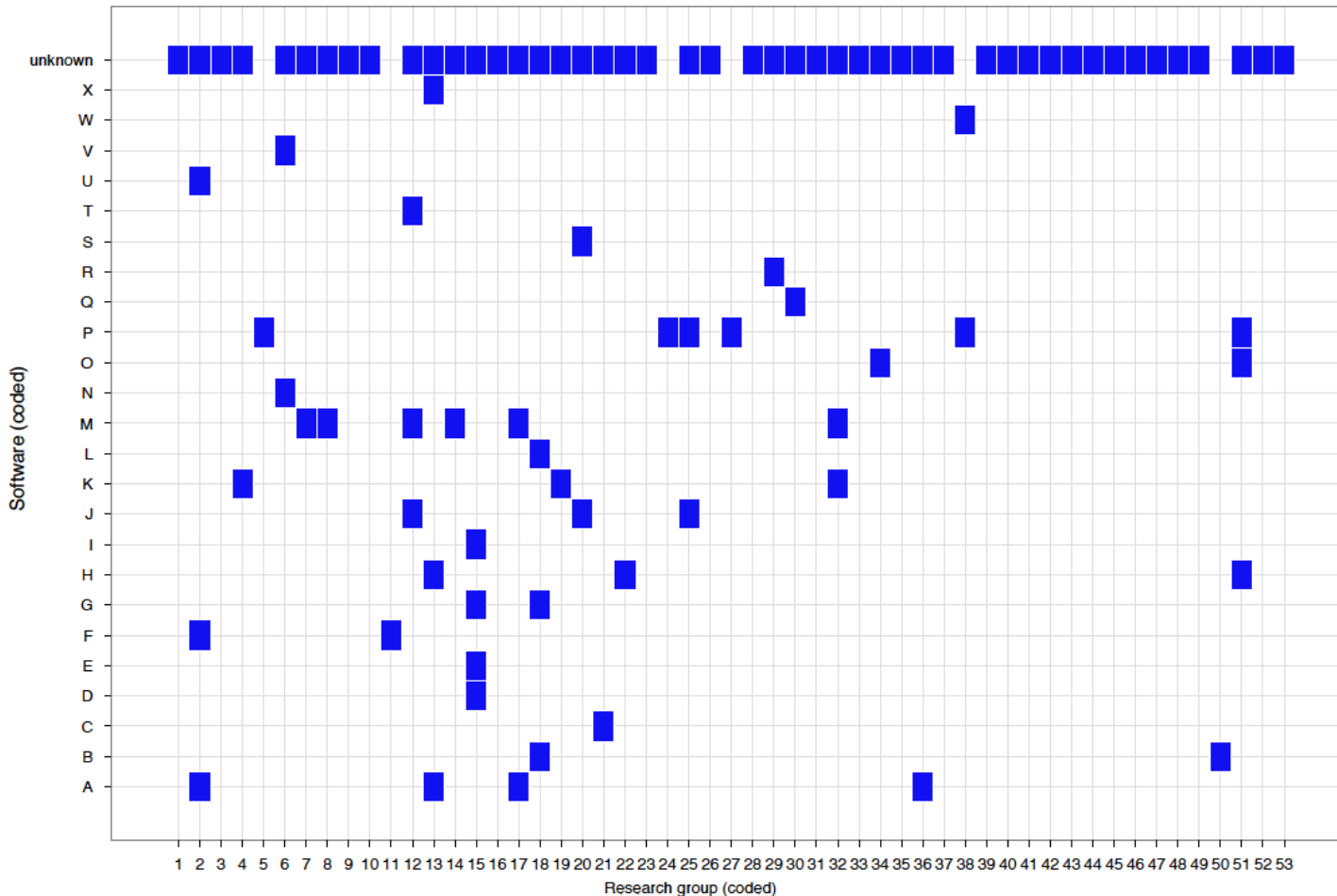


# RDAV - General Data Analysis and Visualization Services

- Located at National Institute of Computational Science
- Goal
  - To provide data analysis and visualization services that enable computation and data intensive science
- Approach
  - Unique HPC system – SMP: 1024 core, 4TB memory
  - Comprehensive expertise



# Diverse Domain Specific Software



## Granularity of Parallelism

Parameter sweep using serial apps: MaxEnt, Energy+ ...

Parallel apps: Paraview, VisIt

Parallel programming: R, Python, Java, Matlab

MPI, pthreads



# Scalable News Mining

Scaling up news mining code in Perl

Used RDAV/UT's in-house developed glue software – Eden (released as rdav\_eden on sourceforge.net)

Quoting from the NCSA user Kalev Leetaru's emails:

.. Scott, this is a *\*fantastic\** tool, exactly what I've been looking for!....

.. I will say that Eden is a great great tool, it is proving *\*tremendously\** useful and really helping to speed things up for me...

The screenshot shows a web browser with two open tabs. The first tab is BBC News - Supercomputer, displaying a BBC News Technology article titled "Supercomputer predicts revolution". The article text states: "Feeding a supercomputer with news stories could help predict major world events, according to US research." It mentions a study based on millions of articles charted the tone of news reports. The second tab is www.nature.com/news/2011/110913/full/news.2011.532.html, displaying a Nature News article titled "News mining might have predicted Arab Spring". The article text states: "Signs of impending social and political change may lie hidden in a sea of news reports." It mentions that computer scientist Kalev Leetaru at the University of Illinois at Urbana-Champaign has trawled through a vast collection of news reporting and examined the 'tone' of the news about Tunisia, Egypt and Libya, where long-established dictatorial regimes have been overthrown. The Nature News article also includes a list of "Stories by subject" (History, Policy, Mathematics) and "Stories by keywords" (Culturomics, Data mining, Conflict, War, Social sciences, News). The article is also linked to other sources like Blogs linking to this article, Add to Connotea, Add to Digg, Add to Facebook, Add to Newsvine, and Add to Delicious.



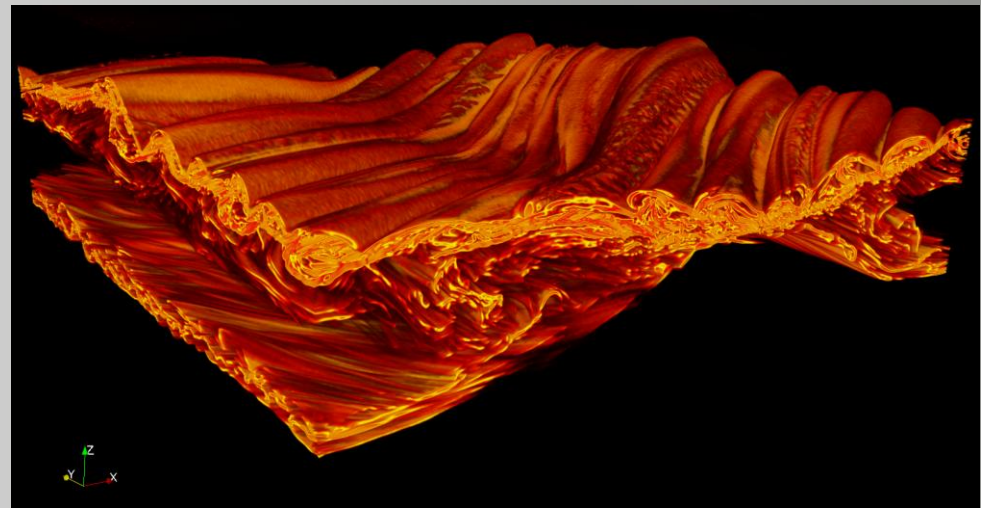
# Visualizing the Magnetosphere

Electron and ion vorticity from the  
asymmetric VPIC simulation

120GB per vector field per timestep.

Visualized using a discrete particle  
noise reduction technique deployed  
in ParaView.

**Collaboration with Homa  
Karimabadi (UCSD/SciberQuest).  
Visualization by RDAV/LBNL.**



# **Internet Engineering Task Force: Open Process for Internet Standards**

Russ Housley  
IETF Chair

Data2012  
January 2012





# Internet Engineering Task Force I E T F<sup>®</sup>

- Formed in 1986
- “We make the net work”
- The mission of the IETF is to produce high quality, relevant technical and engineering documents that influence the way people design, use, and manage the Internet in such a way as to make the Internet work better. These documents include protocol standards, best current practices, and informational documents of various kinds. [RFC 3935]



**I E T F<sup>®</sup>**

# IETF Open Standards

While the mission of the IETF is to make the Internet work better, no one is “in charge” of the Internet. Instead, many people cooperate to make it work. Each person brings a unique perspective of the Internet, and this diversity sometimes makes it difficult to reach consensus. Yet, when consensus is achieved, the outcome is better, clearer, and more strongly supported than the initial position of any participant.



I E T F®

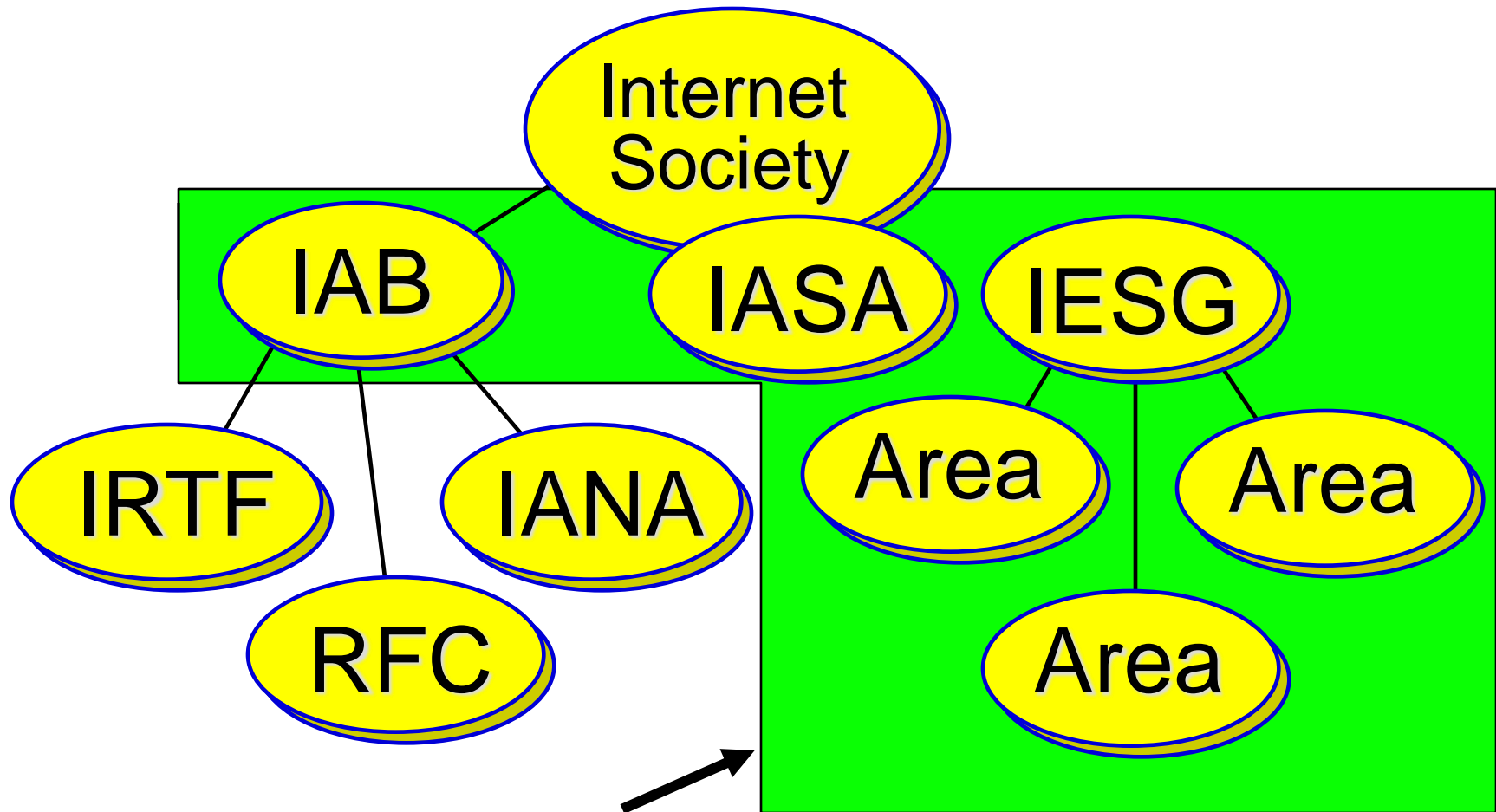
# IETF Structure Overview

- The IETF is not a legal entity – no members
- 1000 to 1200 people at 3/year meetings
  - Many more people on mail lists
- ~120 Working Groups (WGs)
  - Where the real work gets done
- 7 Areas, each lead by two Area Directors (ADs)
  - Except General Area is lead by IETF Chair
- IESG: management, standards approval
- IAB: architectural guidance, liaison, oversight
- IAOC: oversee budget, contracts, and IPR



I E T F®

# IETF Structure Overview



*"The IETF"*

# Ethos of the IETF

- IETF uses an open standards process
  - All interested people are invited to participate
  - Even if unable to attend the face-to-face meetings, invited to join mail list discussions
  - All documents are online, available to everyone
- One Internet
  - Open standards for a global Internet
  - Maximum interoperability and scalability
  - Avoid specialized protocols in different places
- Contributions are judged on merits:  
*rough consensus and running code*





I E T F®

# IETF Moto

*“We reject kings, presidents and voting. We believe in rough consensus and running code.”*

Dave Clark, MIT



# IETF takes on work when ... I E T F<sup>®</sup>

- The problem needs to be solved
- The scope is well defined and understood
- Agreement that the specific deliverables
- Reasonable probability of timely completion
- People willing to do the work

# IETF is right place when ...

- The problem fits one of the IETF Areas
  - Applications
  - Internet
  - Operations and Management
  - Real-time Applications and Infrastructure
  - Routing
  - Security
  - Transport
  - General
- Working on problems that span Standards Development Organizations (SDOs) take *significantly* more effort to be successful



# IETF is successful when ...

**I E T F<sup>®</sup>**

- Participants care about solving the problem
- Participants represent all stakeholders
- Successful Internet protocols have come from top-down and bottom-up approaches
  - Bottom-up is more common today
  - Many efforts are incremental improvements



I E T F<sup>®</sup>

# IETF Management

- **IETF Chair**
  - IESG Chair, AD for General Area, IAB member, IAOC member, also seen as spokesman
- **Area Directors (AD)**
  - Two ADs for each Area other than the General Area
- **Internet Engineering Steering Group (IESG)**
  - ADs sitting as a body
- **Internet Architecture Board**
- IETF Chair, ADs, IAB, and two IAOC members selected by Nominating Committee for 2 year term





I E T F<sup>®</sup>

# Area Directors

- Responsible for setting direction in the Area
- Responsible for managing process in the Area
  - Approve Birds of a Feather (BOF) sessions
  - Appoint working group chairs
  - Oversee working group charters
    - IESG and IAB involved in charter approval
- Review all working group documents prior to IESG evaluation
  - IESG approves all IETF RFCs



I E T F<sup>®</sup>

# IETF Working Groups

- Where the IETF primarily gets work done
  - Most discussion are on mail list
  - Face-to-face meetings focused on key issues
- Working group focused by charter with milestones
- Charter approved by IESG with advice from IAB
- No defined membership – just participants
- “*Rough consensus and running code...*”
  - No formal voting
  - Does not require unanimity
  - Disputes resolved by discussion
  - Final decisions are verified on mail list



I E T F®

# Nominations Committee

- IETF Chair, ADs, IAB and 2 IAOC members are picked by Nominations Committee
  - NomCom Chair appointed by ISOC president
- Volunteer to be a NomCom voting member
  - Must attend 3 of last 5 IETF meetings
  - Ten voting members randomly selected from the volunteer pool
- NomCom picks one person for a 2 year term
- Confirmation before names are announced
  - IETF Chair and ADs confirmed by IAB
  - IAB confirmed by ISOC Board of Trustees
  - IAOC confirmed by IESG

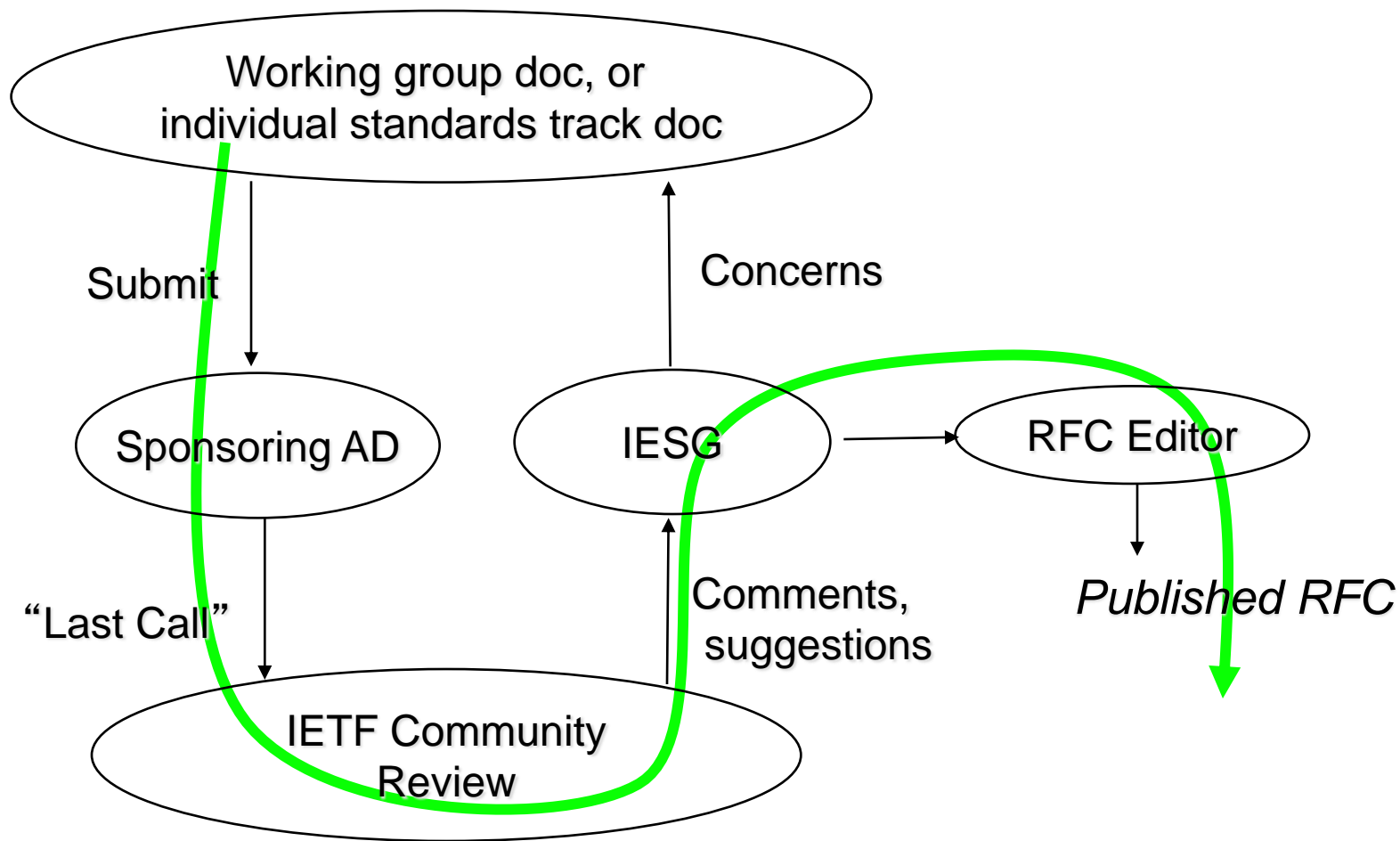


I E T F®

# Voluntary Standards

- The IETF standards enable interoperability only because people choose to use them
  - Some other SDOs can mandate use of their standards
  - No governmental recognition for IETF standards
    - Some indications that this is changing

# IETF Standards Approval





# IETF Summary – IETF Movie



[http://www.youtube.com/watch?v=tqc8vd\\_jPpg](http://www.youtube.com/watch?v=tqc8vd_jPpg)

# Thank You



**I E T F<sup>®</sup>**

**Russ Housley**

Phone: +1 703 435 1775

Email: [housley@vigilsec.com](mailto:housley@vigilsec.com)



**I E T F<sup>®</sup>**

*Engineering the Internet's Future for 25 years.*

**IETF**

Internet Engineering Task Force

[www.ietf.org](http://www.ietf.org)

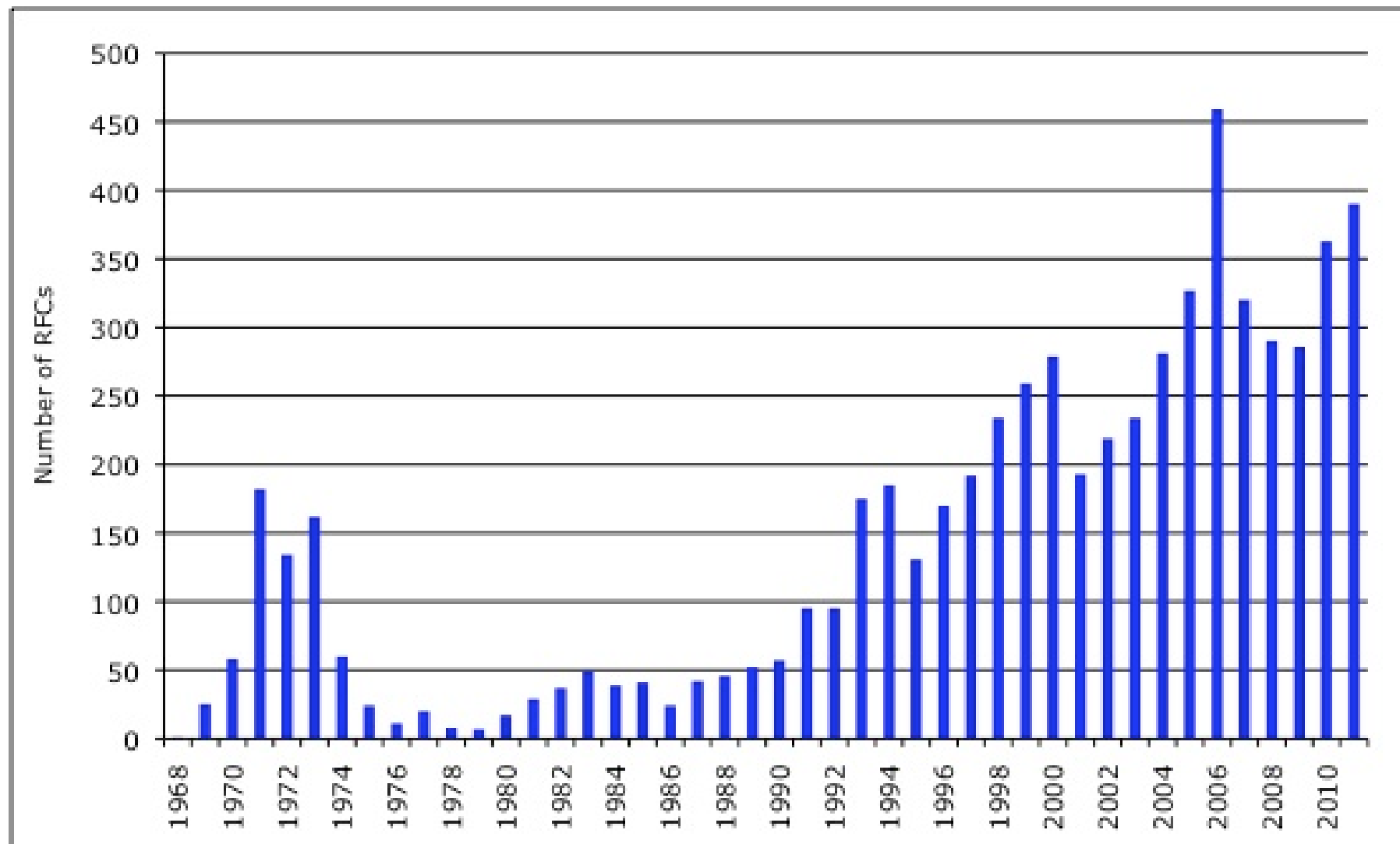


# **Supporting the IETF Standards Development Process**

Ray Pelletier  
IETF Administrative Director

Data2012  
January 2012

# RFCs Published 1969 - 2011







**I E T F<sup>®</sup>**

# Supporting the IETF

History

IASA

Secretariat

Life Cycle Data Tracker

Publication Services

Volunteers

Funding

Expenses

# History

- IETF began 1986
- NSF funded IETF operations from 198X to 1991
- 1992 Internet Society – home of the IETF
  - Funding
  - Insurance
  - Accounting
  - Services
- 2005 IETF Administrative Support Activity (IASA)



**I E T F<sup>®</sup>**

# Supporting the IETF: IASA

- IETF Administrative Support Activity
  - provides the administrative structure
  - IAOC – IETF Administrative Oversight Committee
  - IAD – IETF Administrative Director
    - Only IETF employee
    - Oversee contract support of IETF
- 2012 Budget: \$350k

# Supporting the IETF: Secretariat



- Services
  - Meetings Services – 3 meetings/yr
  - Clerical Services to IESG
  - IT Support
    - Mail lists
    - Website
    - Tools development and support
    - WebEx support
- 2011 Meeting Expenses: \$2.1m
- 2011 Clerical & IT Expenses: \$1.5m

# Supporting the IETF: Life Cycle Data Tracker



- N Internet Drafts introduced or updated in 2011
- Managing the Standards Development Process
  - Invested > \$500k in a re-architecture of data tracker and tools development over two-year period
  - Deploy in 2012
  - Enable monitoring and managing Internet-Drafts from -00 introduction through RFC publication to aid Authors, Working Group Chairs, IETF leadership



# Supporting the IETF: Tools



- N Mailing lists averaging 250 subscribers
- Tools to create Internet Drafts
- Wikis to manage issues raised with a Draft
- Tools and wikis created and maintained by Volunteers

# Supporting the IETF: Publication Services



- RFCs average over 300 per year since 2005
  - Average 30 pages each
- Service contracted out
  - RFC Series Editor
  - RFC Production Center
  - RFC Publisher
  - RSOC – RFC Series Oversight Committee
- 2012 Budget: \$890k

# Supporting the IETF: Volunteers



**I E T F<sup>®</sup>**

- IETF is a Volunteer Organization
  - Internet Engineering Steering Group
  - Working Group Chairs and Members
  - Authors
  - Tool Developers

# Supporting the IETF: Funding



- 2012 Budget: \$5m
- Revenue
  - Meeting Registration: 40%
  - Hosts and Sponsors: 20%
  - Internet Society: 40%

# Thank You



**I E T F<sup>®</sup>**

**Ray Pelletier**

Phone: +1 703 439 2123

Email: [iad@ietf.org](mailto:iad@ietf.org)





**I E T F<sup>®</sup>**

*Engineering the Internet's Future for 25 years.*

**IETF**

Internet Engineering Task Force

[www.ietf.org](http://www.ietf.org)

# How IETF sees work divided

**W3C**

	Mail	HTML	SNMP	Voice/ Video Data	Telephony Signaling
		HTTP			
	TCP		UDP	RTP	
	Internet Protocol				
	Ethernet	ATM	Frame Relay	PPP	MPLS
A variety of physical layers and interfaces					Cellular Radio

**ETSI**

**ITU-T**

**IEEE**



# Technology Work in EUDAT DAITF Idea

Peter Wittenburg, Daan Broeder  
The Language Archive, Max Planck Institute, Netherlands  
DATA2012, Indianapolis



Date: 26<sup>th</sup> January 2012

# Outline of the talk

- ❑ what are the needs of communities wrt common services
  - ❑ which are the enabling technologies
- ❑ Understanding the Collaborative Data Infrastructure
  - ❑ how is data organized in communities
  - ❑ how are data centers organized (not today)
- ❑ Main Service Cases
  - ❑ Safe replication, Data Staging to HPC
  - ❑ Joint Metadata Domain, Simple Data Store
  - ❑ Federation Scalability (volumes, complexity), workflow frameworks
- ❑ Data Access and Interoperability Task Force (DAITF)

# Communities and Data Centers

Which common services are needed?

What are the basic requirements?



# Community Service Wishes

## **In Progress as Services (Task Forces set up)**

- Safe Data Replication (for Bit-stream Preservation & Access Optimization)
- Dynamic Data Replication into HPC Workspace

## **In Specification/Discussion as Services**

- Aggregated EUDAT Metadata Domain
- Researcher Data Store (Simple Upload, Share and Access)

## **In Progress as Research Issues (WP7)**

- more elaborate policy rules and federation scalability
- generic workflow execution framework  
(automatic annotation, data mining, etc.)

All 5 core communities basically share same service wishes  
slightly different priorities



# Enabling Technologies

- **Building robust and available persistent identifier service (is in place based on Handles)**

- EPIC: millions of objects, DataCite: published collections
- EPIC offers registration/resolution service for all data centers in Europe
- EPIC currently 3 strong data centers with redundancy setup
  - will be extended to 10-12 collaborating strong data centers
- EUDAT: all objects need to be registered, all policy operations will use PIDs

ready  
to go

- **Federated AAI service**

- Shib/SAML based world - still a mess due to fragmentation
- can we rely on harmonized EU wide Identity Federation?
- will individual identity providers offer needed attributes?
- do we need to fall back on own user administration?

not yet  
ready

- **Shared Workspaces**

- obviously for different purposes (storing data, automatic annotations, etc)

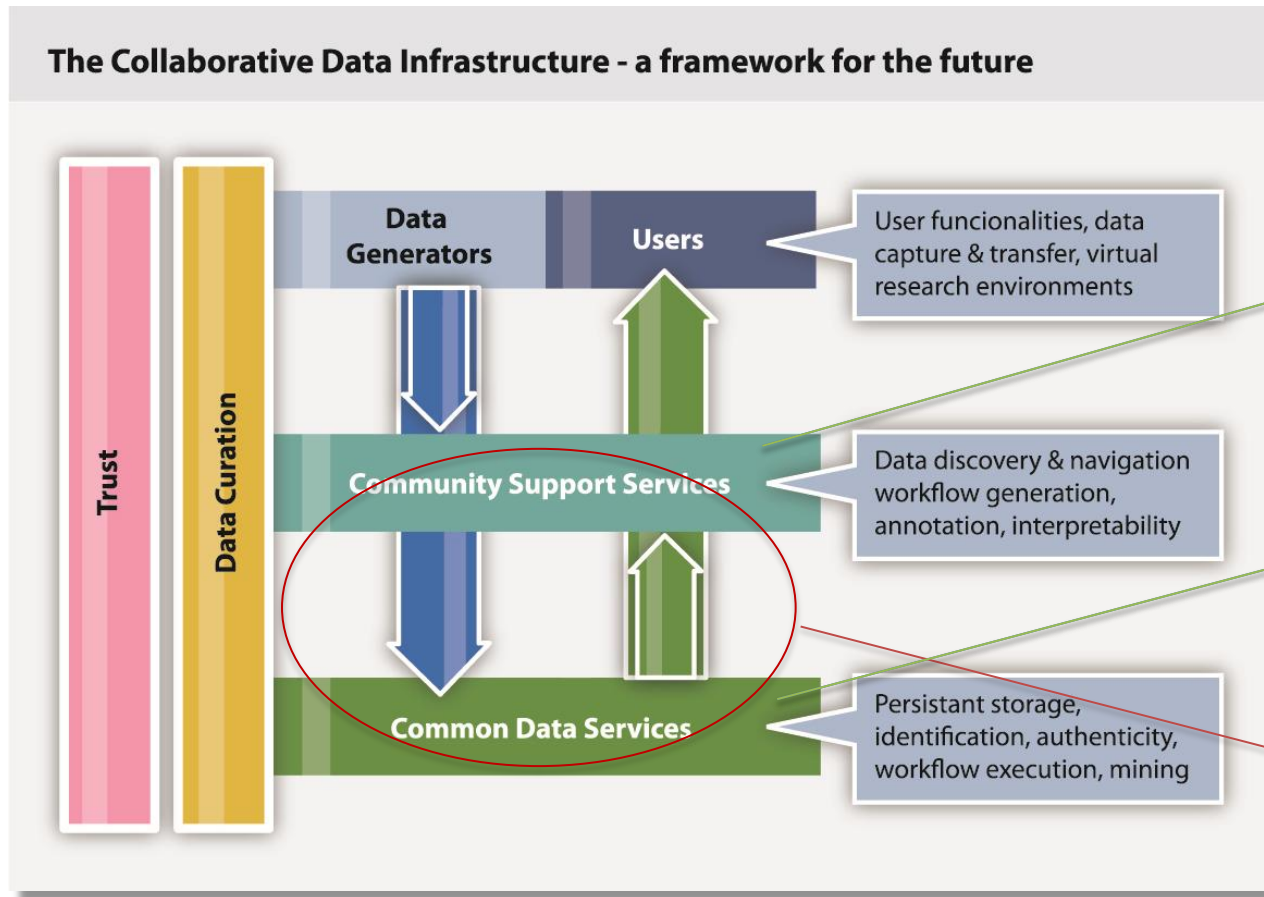
to be  
done

- **Monitoring and accounting**

- all participating servers/services need to show stability, availability

- **Network Services (of course)**

# First - need to understand CDI



CLARIN, LifeWatch, ENES,  
EPOS, VPH, etc.  
5 Core Infrastructures  
~15 second round  
infrastructures

=> 10 EUDAT data centers

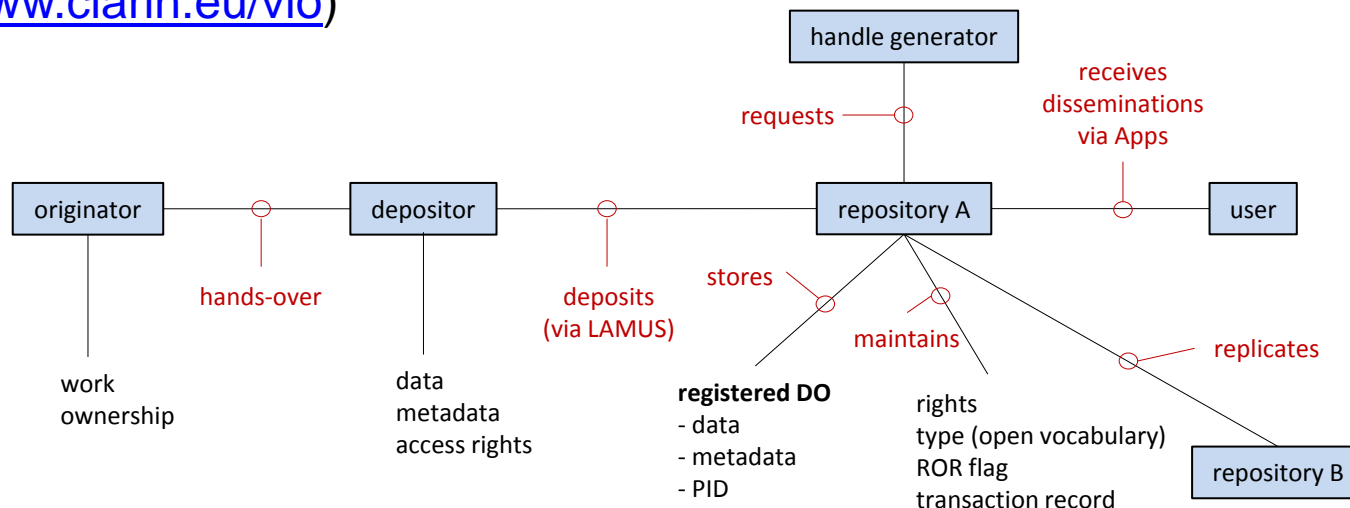
indeed some  
heterogeneity at both  
levels

Interviews based on Abstract Data Object Model  
(interviews fostered architecture clarifications)

# Data Landscape Analysis: CLARIN

- **CLARIN (Language Resource and Technology Community)**

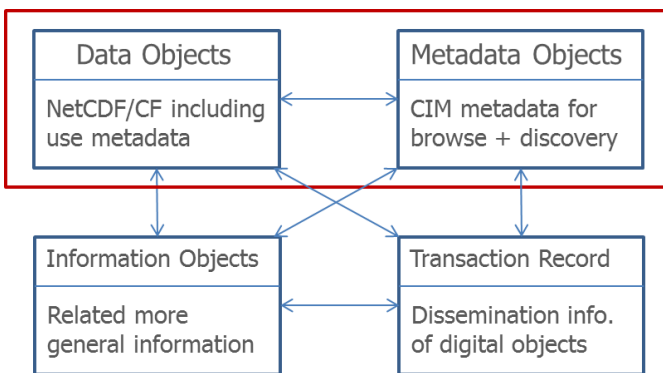
- about 200 centers in Europe with about 30 „community center“ candidates
- have 4 types of centers (DataONE: tiers) from strong to weak requirements
- requirements: rep. system, PIDs, CMDI based metadata, AAI
- almost all busy with re-structuring - only few fulfill strong requirements
- components/profiles and concepts registered (ISOcat, SCHEMcat)
- Virtual Language Observatory: harvesting, mapping, indexing  
([www.clarin.eu/vlo](http://www.clarin.eu/vlo))



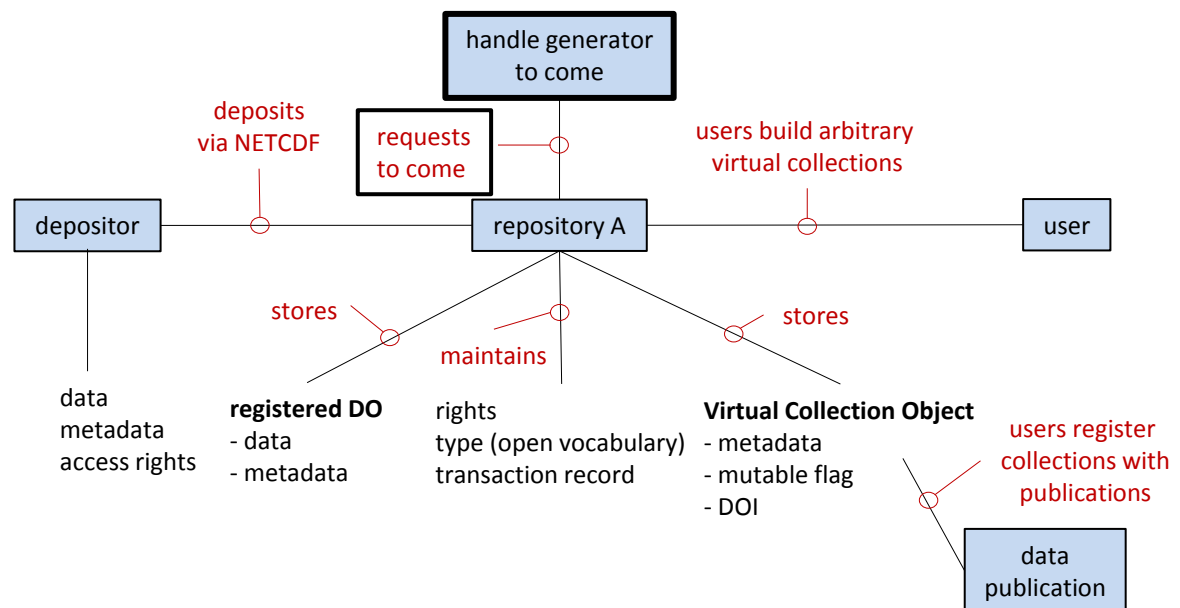
# Data Landscape Analysis: ENES

## • ENES (Climate Modeling Research)

- about 20 centers in Europe -
- have CIM data model - but this is still in a prototype state, not deployed broadly
- but CDI as operating at German Climate Center is taken as basis
- CIM has kind of „canonical“ design using DOIs and EPIC Handles
- Metadata based on ISO 11179 etc.; OAI-PMH in place



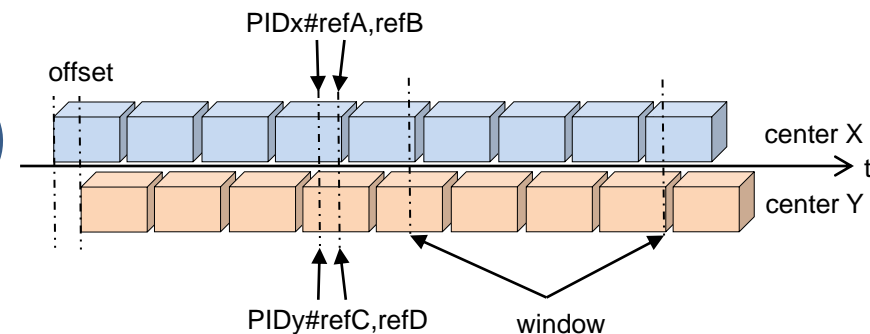
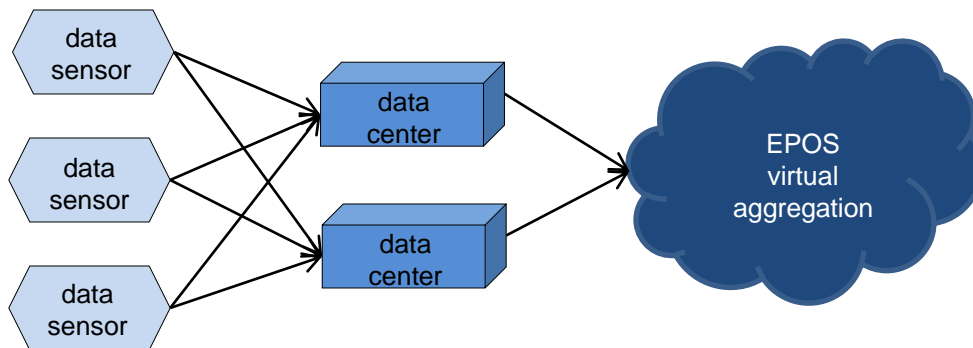
*Identification of distinct data objects and P2P infrastructure*



# Data Landscape Analysis: EPOS

- **EPOS (Seismologists, Vulcanologists, etc.)**

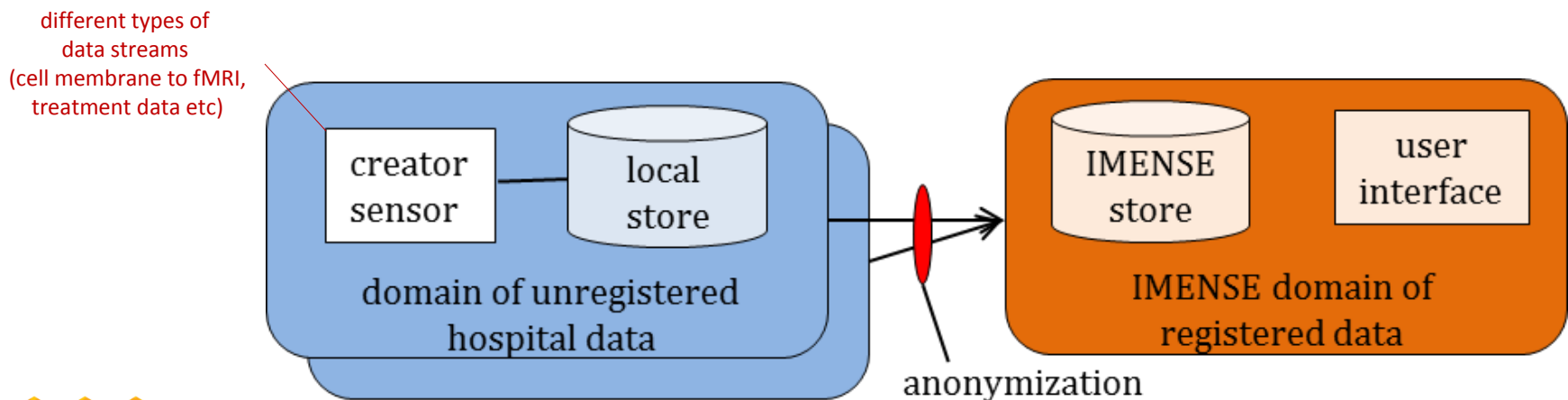
- lots of distributed data sensors producing continuous package streams
- due to various reasons data streams include gaps to be filled over time
- data windows of interest (Wol) are defined „vulcano eruption X“
- aggregations of such data are of relevance (large scale statistics etc)
- work currently on a description of metadata schema for Wols
- work on a scheme of how to refer to packages and offsets (Handles, fragments)
- one center is now implementing reference architecture
- need to synchronize with US and other colleagues



# Data Landscape Analysis: VPH

- **VPH (Virtual Physiology of Humans)**

- currently pilot project with about 5 hospitals in different countries
- one centralized data center - in next phase distributed system
- focus was on metadata aggregation
- IMENSE stores all textual data and Metadata in a DBMS and gives access
- data aggregation is planned together with a large data center in EUDAT
- metadata not yet standardized & formalized (DICOM, JPEG headers, etc.)
- nothing done with PIDs, AAI and OAI-PMH yet







# Data Landscape Analysis: LifeWatch

- **Biodiversity (much based on GBIF)**

- yet no chance of qualified interaction due to time restrictions
- different contributors and actors
- very heterogeneous domain
- first requirements & implementations without LifeWatch
- need to be flexible enough anyhow

# Data Landscape Analysis: 2nd Round

- second round of interviews to come in February/March
- User Forum (March) to meet even other initiatives and start interactions

Environmental Science	ENES, EPOS, Lifewatch, EMSO, IAGOS-ERI, ICOS, Euro-Argo, ...
Social Sciences and Humanities	CLARIN, CESSDA, DARIAH, ...
Biological and Medical Science	VPH, ELIXIR, BBRMI, ECRIN, DiXA, ...
Physical Sciences and Engineering	WLCG, ISIS, DESY, PanData, ...
Material Science	ESS, ...



# Data Landscape Analysis: Summary

- **panta rei - all is moving**

- data infrastructures are shooting on a moving target
  - from core communities only 2 have a ready made architecture
- process of discussion is rather fruitful
  - forces explicitness and fosters harmonization
  - discussions and moderation roles are highly appreciated
- data volumes ready to be contributed range from Exabytes to Terabytes

# Back to community Service Wishes

## **In Progress as Services (Task Forces)**

- Safe Data Replication (for Bit-stream Preservation & Access Optimisation)
- Dynamic Data Replication into HPC Workspace

## **In Specification/Discussion as Services**

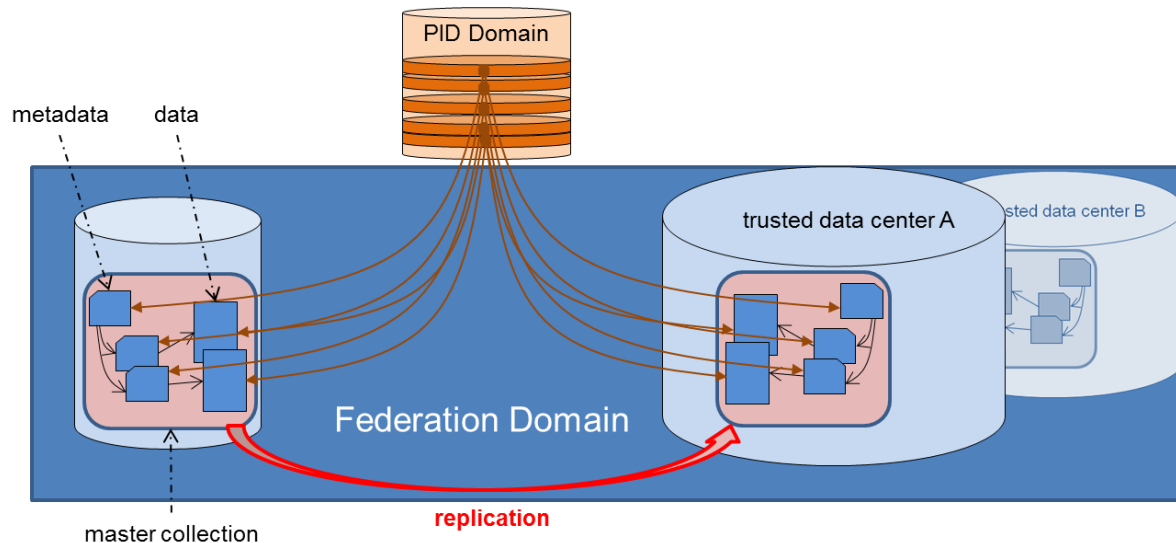
- Aggregated EUDAT Metadata Domain
- Researcher Data Store (Simple Upload, Share and Access)

## **In Progress as Research Issues (WP7)**

- more elaborate policy rules and federation scalability
- generic workflow execution framework  
(automatic annotation, data mining, etc.)

# SAFE Data Replication

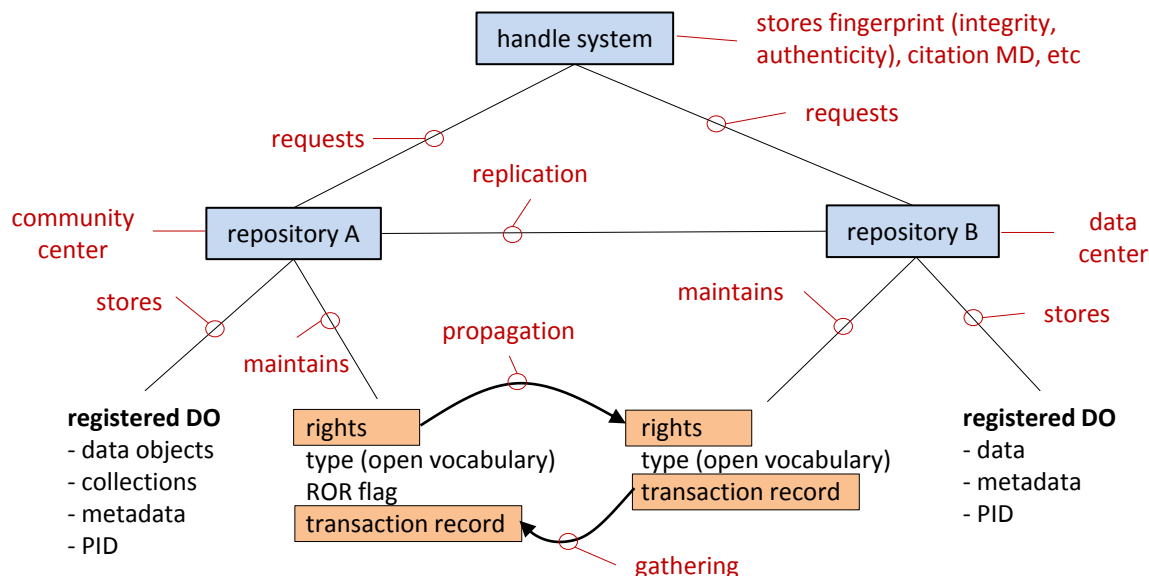
- safe replication between 1 community center and N data centers
- flexibility, scalability and management require policy rule based approach
- 3 islands (community + data center) in parallel & close interaction



- basic technologies: AAI, iRODS, Handles, community MD & OAI-PMH, center registry
- in June merging of 3 islands to one flexible replication domain
- REPLIX experience is basis

# REPLIX

- safe replication between CLARIN center and RZG data center
- purpose: preservation, computation (AV Recognition) and access optimization
- total amount: 80 Terabytes
- requires policy rule based approach due to quality assessment (Data Seal)
- iRODS, Handles, CMDI Metadata
- deployment of Archive/Access software stack as well



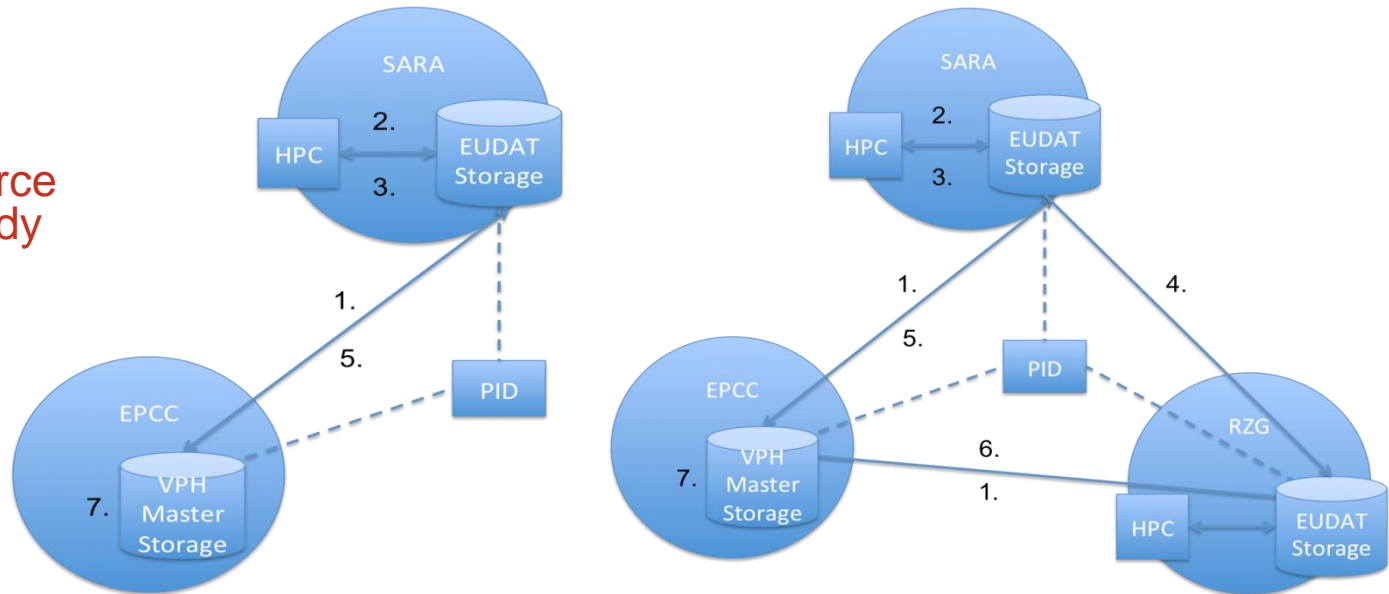
replication at logical collection level basis for demos at ASIST and ICRI conferences both in March (MPI - RENCI)



# Staging to HPC Pipes

- intention is to make use of HPC machines for computations on stored data
- different configurations possible:
  - computations on a single HPC node where data already is
  - computations on multiple nodes - use of PRACE fast distributed file system

Expert Task Force  
built, to be ready  
in summer



- principles:
  - user issues a compute command
  - script pushes data into the HPC workspace, results go into workspace
  - input data is discarded after job end, user needs to store the results

# Aggregated Metadata Domain

- not yet fully specified
- question: for what ???
  - probably loss of specific information - thus interdisciplinary research
  - should show what is stored in the EUDAT data centers
  - one stop shop for virtual collection building
  - making PR for collections (ANDS model)
- general index with some faceted browsing machine probably not sufficient
  - element semantics probably too different
- therefore currently analysis of semantics and simple mapping schemes
- enabling technologies:
  - OAI-PMH, refs via PIDs, SOLR/Lucene for indexing/browsing
  - when and how semantic expansion
  - do we need higher performance technology?
- decision about criteria in February
- technology watch in March

# Researchers Simple Store

- not yet fully specified
- question: for what ???
  - researchers need/want Simple Store for all their „secondary“ data
  - trust is an important issue - owner/copyright must be (with) the researcher
  - data should be part of the EUDAT data domain (thus Metadata, PIDs)
  - ingest via community control to prevent misuse
- Simple Store must have simple access component (like YouTube) and perhaps easy ‚promotion‘ of data into community center collections
- enabling technologies:
  - AAI, PIDs, MD Indexing
- decision about criteria in February
- technology watch in April (what about Mercury etc.)

# EUDAT CDI Summary

- understand data organizations as bottom-up exercise
- determine „common“ functions needed
- determine essential independent components with chance of wide acceptance
  - PID system, center registry, metadata landscape
- define agreed APIs for different components
- rely on policy-rule based approach
- currently implementation of procedures for 3 islands
- probably need to extract common characteristics to scale up
- are looking for close collaboration with others (US, etc.)

# What about DAITF?

- Do we need a **Data Interoperability and Access Task Force**?
- We have already:
  - IETF, W3C, OASIS, OAI, CODATA, GRDI, e-IRG,
  - ISO, ISO/IEC JTC1, ITU, MOIMS (RAC), DSA, etc.
- Many promising initiatives world-wide dealing with the same questions
  - data is global, communities are acting globally
  - much overlap in intentions - however slight differences
- Our conclusion: we need a forum (whatever name we give it) where
  - “data practitioners” can meet regularly - no PR, no politics
  - we can exchange approaches & technologies, discuss harmonization, standards, IT principles, etc.
  - we can train young “data scientists”
- Are we already strong enough to go outside?

# What about DAITF?

- who: data architects, data practitioners, information experts, ?
- what first:
  - need to define the scope
    - there is so much community specific “pre-registration” activity
  - need to meet with a prepared agenda
  - need to have a start-up Steering Group and perhaps first WGs
- agenda:
  - first discussion at DAITF Preparation Workshop at ICRI in Copenhagen - March 20/21.
  - EUDAT/OpenAIRE received money to host 2 Workshops 12/13
  - submitted an application to EC with DAITF continuation
- for EC Data Infrastructures will be a top priority in Horizon 2020
  - EC is going to continue funding DAITF